Loss and Risk in Smoothing Parameter Selection

By

Birgit Grund, Dept. of Applied Statistics, University of Minnesota
Peter Hall, Australian National University[1]
J.S. Marron, Université Catholique de Louvain[2]

School of Statistics, University of Minnesota
Technical Report No. 584

November, 1992

---

[1] Australian National University; CSIRO Division of Mathematics and Statistics.

[2] Université Catholique de Louvain; Limburgs University Center; University of North Carolina, Chapel Hill. Research partly supported by NSF Grant No. DMS-9203135.

November 1992

LOSS AND RISK IN SMOOTHING PARAMETER SELECTION

Birgit Grund[1] Peter Hall[2] J. S. Marron[3]

CMA–SR00–91

AMS (1985) SUBJECT CLASSIFICATION. Primary 62G05, Secondary 62C05.

ABSTRACT. For several years there has been debate over the relative merits of loss and risk as measures of the performance of nonparametric density estimators. In the way that this debate has dealt with risk, it has largely ignored the fact that any practical bandwidth selection rule must produce a random bandwidth. Existing theory for risk of density estimators is almost invariably concerned with nonrandom bandwidths. In the present paper we examine two different definitions of risk, both of them appropriate to circumstances where the bandwidth is random. Arguments in favor of, and motivations for, each approach are presented, including formulation of appropriate decision-theoretic frameworks. It is shown that the two approaches can give diametrically opposite answers to the question of which of two competing bandwidth selection rules is superior. Technical results include some surprising conclusions about the nonexistence of risks, and even of moments of some common data-driven bandwidths under the usual assumptions.

KEY WORDS AND PHRASES. Bandwidth selection, decision theory, density estimation, kernel, loss, risk, smoothing parameter.

SHORT TITLE. Smoothing parameter selection

[1] University of Minnesota.

[2] Australian National University; CSIRO Division of Mathematics and Statistics.

[3] Université Catholique de Louvain; Limburgs University Center; University of North Carolina, Chapel Hill. Research partly supported by NSF Grant No. DMS-9203135.

LOSS AND RISK IN SMOOTHING PARAMETER SELECTION

Birgit Grund[1] Peter Hall[2] J. S. Marron[3]

1. Introduction

Many important ideas in nonparametric curve estimation have been developed in terms of kernel density estimators, because of their simplicity of both analysis and presentation. Based on a simple random sample $X_1, \ldots, X_n$ from an unknown probability density f, a kernel estimator of $f(x)$ is given by

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} . \tag{1.1}$$

Here, $h > 0$ denotes the bandwidth and K, the kernel function, satisfies $\int K = 1$. Throughout this paper we confine attention to the case where K is a symmetric density function. More general, higher-order kernels may be treated similarly, albeit with more complex notation. See Silverman (1986, Chapter 3) for discussion of important aspects of this estimator.

The performance of $\hat{f}_h$ is governed largely by choice of the smoothing parameter, h. If h is too small then $\hat{f}_h$ tracks the data too closely, and is too "wiggly"; if h is too large then important underlying features of f may be "smoothed away" by $\hat{f}_h$, and ignored. A regular and theoretically tractable measure of error is the Integrated Squared Error, or ISE, defined by

$$\Delta(h) = \int (\hat{f}_h - f)^2 . \tag{1.2}$$

[1] University of Minnesota.

[2] Australian National University; CSIRO Division of Mathematics and Statistics.

[3] Université Catholique de Louvain; Limburgs University Center; University of North Carolina, Chapel Hill.

The ISE measures how closely the kernel estimate fits the true density for *the particular sample at hand.* Therefore, it is common in kernel smoothing to view ISE as "loss". Historically, the bandwidth has been viewed as nonrandom in theoretical investigations, so the Mean Integrated Squared Error, or MISE, naturally represents "risk". However, if the bandwidth is viewed as random (for example in data-based bandwidth selection), MISE, evaluated at the random bandwidth, is no longer the same as the expected ISE. Nevertheless, the term "risk" has been used for both, often casually ignoring the strict relation of "loss" and "risk" in the decision-theoretic sense.

Much work has been done on data-based bandwidth choice, resulting in an overwhelming variety of methods, see Marron (1988), Cao-Abad, Cuevas and Gonzalez-Mantiega (1992), Park and Turlach (1992) and Jones, Marron and Sheather (1992) for an overview. While the methods may vary, their motivations may be grouped into two essential types, according to the viewpoint on which they are based.

Viewpoint V_μ: The classical viewpoint concentrates on Mean Integrated Squared Error with a nonrandom bandwidth, h,

$$M(h) = E\int(\hat{f}_h - f)^2\,. \tag{1.3}$$

Clearly, in kernel density estimation with deterministic bandwidth, $M(h) = E\{\Delta(h)\}$ represents risk, defined as expected loss. Under viewpoint V_μ, the optimal smoothing parameter is h_μ, one of the *deterministic* minimizers of $M(h)$; the method of breaking ties is not important, here and elsewhere.

Random bandwidths are studied mathematically through the random quantity $M(\hat{h})$, obtained by simply substituting $\hat{h}$ for h in (1.3), while ignoring the data-dependence of $\hat{h}$ in taking the expectation. In this situation, it is not possible to view $M(\hat{h})$ as a measure of risk, but the expected value of that quantity,

$$\mu(\hat{h}) = E\{M(\hat{h})\}\,,$$

does admit this interpretation. Later in this section we shall provide a decision-theoretic basis for defining risk in terms of $\mu(\hat{h})$.

Note that $\mu(\hat{h})$ is minimized by a deterministic bandwidth. It is again the same optimal parameter h_μ that minimizes $M(h)$. Hence, the viewpoint V_μ finally aims at estimating the constant h_μ as well as possible.

Viewpoint V_ν: There is no difference between V_μ and V_ν in respect of nonrandom bandwidths. In the case of *random* bandwidths, however, advocators of V_ν use the ISE $\Delta(\hat{h})$ as "loss" in the correct decision-theoretic sense. The overall behavior of the density estimator, including random bandwidth, is measured by Expected Integrated Squared Error, or EISE,

$$\nu(\hat{h}) = E\{\Delta(\hat{h})\} = E\int(\hat{f}_{\hat{h}} - f)^2 . \tag{1.4}$$

In taking the expectation, the location of the kernels and the random bandwidth are treated *simultaneously* as data-dependent.

The risk $\nu(\hat{h})$ reflects the fact that *each* sample imports randomness to $\hat{f}_{\hat{h}}$ through two components: the location of the n kernels, that are centered at $X_1,\ldots,X_n$, respectively, and the scale of the kernels, determined by the data-dependent bandwidth $\hat{h}$. A good visualization of these two components is provided in Figures 2.4 and 2.5 of Silverman (1986).

Let us denote by $\hat{h}_\nu$ the random bandwidth that minimizes $\Delta(\hat{h})$ pointwise. Obviously, $\hat{h}_\nu$ is an optimal smoothing parameter with respect to risk ν, as it minimizes $\nu(\hat{h}) = E\{\Delta(\hat{h})\}$ as well. Thus, the viewpoint V_ν indirectly aims at estimating the random quantity $\hat{h}_\nu$.

Both $\mu(\hat{h})$ and $\nu(\hat{h})$ have been studied numerically by other authors, although little attention has been paid to their roles as measures of risk, see Hall and Marron (1991) for a review. Instead, they have emerged in a natural way in Monte Carlo studies, with $\mu(\hat{h})$ and $\nu(\hat{h})$ being effectively approximated by averages of $M(\hat{h})$ and $\Delta(\hat{h})$, respectively, over sequences of Monte Carlo trials. One of the

main contribution of the present paper is to provide the first proper mathematical treatment of μ and ν. Earlier work of a variety of authors focussed only on limiting distributions for $M(\hat{h})$ and $\Delta(\hat{h})$. Statistical folklore has been that it should not make much difference whether performance of a bandwidth selector is measured by μ or ν. However, we show that much more care is needed on this point, since there are two important ways of selecting $\hat{h}$ which have diametrically opposite relative performance with respect to μ and ν.

This makes the question of which viewpoint is preferred more important than ever. However, already there has been considerable controversy on this topic; see for example Mammen (1990) and Jones (1991). We provide two new insights into this debate. Firstly, we show that both μ and ν have interpretations in terms of the concept of classical "risk" from decision theory, although based on quite different loss functions. Secondly, we present the randomness of the final density estimator $\hat{f}_{\hat{h}}$ as consisting of two parts, which may be described roughly as "location of the kernel centerpoints" and "scale of the kernel functions". The differences between μ and ν lie in the way these measures of risk deal with the two different sources of randomness. Separating location and scale provides a useful intuitive basis for comparing both approaches. Finally, we summarize other arguments and motivations which have been made in favor of μ and ν.

Decision-theoretic interpretation. There is considerable disagreement as to whether μ or ν should be used in assessing the performance of a bandwidth. Nevertheless, both sides of this debate tend to accept the ISE, $\Delta(h)$, as reasonable distance for measuring departure of the density estimator from the true density "pointwise", *for the data at hand.*

The main argument supporting ν is that $\nu = E\{\Delta(\hat{h})\}$ is the decision–theoretic proper risk corresponding to loss $\Delta(h)$. Decision problems are well–defined by the triple action space $\mathcal{A}$, parameter space Θ, and loss function $L : \mathcal{A} \times \Theta \longrightarrow I\!R$. Let

us investigate the decision–theoretic background for both viewpoints V_μ and V_ν. We shall see that both μ and ν can be interpreted as "risks", corresponding to different underlying decision problems.

V_ν: The representation of ν as risk is straightforward. The underlying decision problem is, shortly speaking, "given the density f, choose the best possible kernel estimator".

Let $\mathcal{X}$ denote the sample space. Then the action space $\mathcal{A}_\nu$ is the set of all functions

$$\mathcal{A}_\nu = \{f_h : f_h(x|x_1,\ldots,x_n) = (nh)^{-1}\sum_{i=1}^{n} K\{(x-x_i)/h\}\}$$

for certain values $x_1,\ldots,x_n \in \mathcal{X}$, i.e. the action space consists of all curves that can be represented as realizations of kernel estimators. The parameter space Θ_ν is the set of all densities (restricted by regularity conditions, if necessary). We consider the kernel estimators $\hat{f}_{\hat{h}}$ as decision functions, obtained by substituting h in (1.1) by the data-driven bandwidth $\hat{h} = \hat{h}(X_1,\ldots,X_n)$. Then, with "loss" $L_\nu(f_h, f) = \int(f_h - f)^2$, we obtain the "risk"

$$\nu(\hat{h}) = E\int(\hat{f}_{\hat{h}} - f)^2 = E\{\Delta(\hat{h})\}. \tag{1.5}$$

Note that each sample defines its own data-dependent bandwidth, and that the expectation averages simultaneously over both sources of variability: the location of the kernels and their scale. The optimal decision function is the kernel estimator that minimizes (1.5). In this context, the optimal smoothing parameter $\hat{h}_\nu$ would be the random bandwidth corresponding to the optimal kernel estimator.

V_μ: The classical measure μ corresponds to a different setting. Here, the focus is on the optimal bandwidth rather than an optimal curve estimator. At this stage, we do not take into account the relationship among the estimator, the bandwidth and the data at hand.

The parameter space is again the set of all densities , but the action space under viewpoint V_μ is now the set of all bandwidths $\mathcal{A}_\mu = \{h : h > 0\}$. Hence, the decision functions to be compared by the measure of risk are the random bandwidths $\hat{h} = \hat{h}(X_1, \ldots, X_n)$.

We define loss by

$$L_\mu(h, f) = E \int (\hat{f}_h - f)^2 = M(h), \tag{1.6}$$

where h is taken to be nonrandom. Note that the loss averages over all possible locations of the n kernels, for fixed h. The corresponding risk has the formulation

$$\mu(\hat{h}) = E\{M(\hat{h})\} = E\{L_\mu(\hat{h}, f)\}.$$

In this setting, the operation of averaging over the data in their role of determining the locations of the kernels is treated separately from that of averaging over the random bandwidths. It is easy to see that the risk $\mu(\hat{h})$ is minimized by h_μ, the deterministic (rather than random) bandwidth that minimizes (1.6). Thus, under viewpoint V_μ the optimal random bandwidth is chosen to minimize squared error loss for an "average" data set, averaged with respect to location of the kernel centers.

The debate over relative merits of integrated squared error and mean integrated squared error has been joined by Hall and Marron (1987a, 1991), Mammen (1990), Jones (1991), Hall and Johnstone (1992), and discussants of the latter paper. References to many earlier works, that simply used V_μ or V_ν without justification, are given in Hall and Marron (1991). We believe the present paper demonstrates that this controversy is ultimately a matter of personal taste in choice of a loss function; it is not possible for any side to convince the other. Discussions with colleagues gave us the impression that most statisticians prefer V_ν, because the loss L_ν is appealing, despite of the technical challenge of minimizing ν.

Nevertheless, there are statisticians that prefer V_μ. This amounts to a separation of the variability due to location of kernel centers and kernel widths, together

with an averaging over the locations. So L_μ can be viewed as working with "average noise" instead of "realizationwise noise", which has another type of appeal. Besides this, the risk μ is technically easier to minimize than ν.

However, our simulations indicate that the above controversy is mostly of philosophical interest. For a given bandwidth selector, the μ and ν assessments of its performance were quite similar. An interesting fact is that the methods based on V_μ were mostly better than the V_ν methods, in *both* of the senses μ and ν. In particular, some of the V_μ methods were closer to the optimal ν value than those methods which explicitly target ν. Asymptotic analysis indicates otherwise, and we conclude that very large samples are needed for the asymptotics to be effective.

Section 2 discusses a variety of practical rules for computing empirical bandwidths $\hat{h}$, including cross-validation and plug-in rules. The performance of each of these methods, measured by $\mu(\hat{h})$ and $\nu(\hat{h})$, is described in Section 3. Section 4 presents a simulation study that investigates our asymptotic conclusions for finite data sets, and Section 5 contains a proof of the theorem from Section 3.

To our knowledge, this paper provides the first analysis of the asymptotic properties of either $\mu(\hat{h})$ and $\nu(\hat{h})$. There have been several studies of integrated squared error, $\Delta(\hat{h})$, starting with Hall and Marron (1987a, 1987c), and the mean integrated squared error, $M(\hat{h})$, has been studied widely, but we are not aware of attempts to theoretically investigate the corresponding risks.

2. Data based bandwidth selectors. First let us recall some facts about the optimal bandwidths h_μ and $\hat{h}_\nu$ that minimize M and Δ, and therefore μ and ν, respectively. Since $M(h) \sim c_1(nh)^{-1} + c_2 h^4$, where $c_1 = \int K^2$ and $c_2 = \frac{1}{4}\{\int z^2 K(z)\,dz\}^2 \int (f'')^2$, then $h_\mu \sim c_0 n^{-\frac{1}{5}}$, where $c_0 = (c_1/4c_2)^{\frac{1}{5}}$. The value of $M''(h_\mu) \sim c_3 n^{-\frac{2}{5}}$, where $c_3 = 2c_1 c_0^{-3} + 12 c_2 c_0^2$, is of importance to the work in Section 3. It may be proved (see Hall and Marron 1987a) that $\hat{h}_\nu / h_\mu \to 1$, and in fact $n^{\frac{3}{10}}(\hat{h}_\nu - h_\mu)$ is asymptotically normal $N(0, \sigma^2_{\mu,\nu})$, where $\sigma^2_{\mu,\nu} > 0$ is

defined in Section 3.

We shall consider four practical, data-driven versions of $\hat{h}$, of which perhaps the best-known is that defined by cross-validation. Here, $\hat{h} = \hat{h}_{cv}$ is chosen to minimize

$$\mathrm{CV}(h) = \int \hat{f}_h(x)^2\,dx - 2\{n(n-1)\}^{-1} \sum\sum_{i \neq j} K\{(X_i - X_j)/h\}\,,$$

which represents an empirical approximation to $\Delta(h) - \int f^2$.

Other versions of $\hat{h}$ may be calculated using the so-called "plug-in method", and are based on the fact that

$$h_\mu \sim c_0\, n^{-\frac{1}{5}} = \kappa_1 (n J_2)^{-\frac{1}{5}}\,, \tag{2.1}$$

where $J_j = \int (f^{(j)})^2$ and $\kappa_1 = [(\int K^2)/\{\int z^2\, K(z)\,dz\}^2]^{\frac{1}{5}}$. Of course, κ_1 is known, and so only J_2 in formula (2.1) is unknown. As a prelude to estimating J_2, assume that f has at least four bounded derivatives, and that

$$\begin{aligned} &\text{for some } \tfrac{1}{4} < \eta \leq 1,\ c > \tfrac{1}{2} \text{ and } C > 0, \\ &|f^{(4)}(x+y) - f^{(4)}(x)| \leq C|y|^\eta (1+|x|)^{-c} \end{aligned} \tag{2.2}$$

for all $-\infty < x, y < \infty$. Then a kernel estimator of J_2, $\hat{J}_2$ say, may be constructed with the properties $n^{\frac{1}{2}}(\hat{J}_2 - J_2) \to N(0, \tau_2^2)$ in distribution, $nE(\hat{J}_2 - J_2)^2 \to \tau_2^2$ and $P(|\hat{J}_2 - J_2| > n^{-\frac{1}{2}+\epsilon}) = O(n^{-\lambda})$ for all $\epsilon, \lambda > 0$, where $\tau_2^2 = 4\{\int (f^{(4)})^2 f - J_2^2\}$. See Hall and Marron (1987b, 1991). In view of (2.2), a simple plug-in estimator of $\hat{h}_\nu$ is given by

$$\hat{h}_{pi,1} = \kappa_1 (n \hat{J}_2)^{-\frac{1}{5}}\,. \tag{2.3}$$

While $\hat{J}_2$ is root-n consistent for J_2, the ratio $\hat{h}_{pi,1}/h_\mu$ does not converge to unity at rate $n^{-\frac{1}{2}}$, since a significant remainder term has been omitted from formula (2.1). A more concise approximation to h_μ is given by

$$h_\mu = \kappa_1 (n J_2)^{-\frac{1}{5}} + \kappa_2\, J_2^{-\frac{3}{5}}\, J_3\, n^{-\frac{3}{5}} + O(n^{-\frac{4}{5}})\,,$$

where $\kappa_2 = \frac{1}{20}(\int K^2)^{\frac{3}{5}} \{\int z^2 K(z)\,dz\}^{-\frac{1}{5}} \int z^4 K(z)\,dz$. Assuming (2.1), a kernel estimator $\hat{J}_3$ of J_3 may be constructed with the property that for some $\epsilon > 0$ and all $\lambda > 0$, $P(|\hat{J}_3 - J_3| > n^{-\frac{1}{10}-\epsilon}) = O(n^{-\lambda})$. The reader is referred to Hall *et al.* (1991) and to Jones and Sheather (1991) for details. Defining

$$\hat{h}_{pi,2} = \kappa_1 (n\hat{J}_2)^{-\frac{1}{5}} + \kappa_2\, \hat{J}_2^{-\frac{3}{5}}\, \hat{J}_3\, n^{-\frac{3}{5}}\,, \tag{2.4}$$

we obtain a root-n consistent bandwidth selection rule for h_μ; i. e. $\hat{h}_{pi,2}$ has the property $(\hat{h}_{pi,2} - h_\mu)/h_\mu = O_p(n^{-\frac{1}{2}})$.

Depending on the variants of $\hat{J}_2$ and $\hat{J}_3$ used in formulae (2.3) and (2.4), either quantity can be negative. The manner in which those estimators are modified to take account of this difficulty can affect even the finiteness of $\mu(\hat{h})$ and $\nu(\hat{h})$. This matter will be treated in Section 3.

Our final practical bandwidth selection rule is based on a suggestion of Hall and Johnstone (1992) for producing a bandwidth which is asymptotically optimal in the sense of providing the best data-driven approximation to $\hat{h}_\nu$, as distinct from h_μ. To this end, let $\hat{J}_1$ denote a kernel estimator of J_1 (see Hall and Marron 1987b) with the property that $n^{\frac{1}{2}}(\hat{J}_1 - J_1) \to N(0, \tau_1^2)$ in distribution, $nE(\hat{J}_1 - J_1)^2 \to \tau_1^2$ and $P(|\hat{J}_1 - J_1| > n^{-\frac{1}{2}+\epsilon}) = O(n^{-\lambda})$ for all $\epsilon, \lambda > 0$, where $\tau_1^2 = 4\{\int (f'')^2 f - J_1^2\}$. Let $\hat{J}_2$ be as defined earlier, assume K has a bounded derivative, and assume $\hat{h}_{\hat{\mu}}$ is a root-n consistent estimator for h_μ, e. g. $\hat{h}_{\hat{\mu}} = \hat{h}_{pi,2}$. Put $k_1 = \int K^2$, $k_2 = \int z^2 K(z)\,dz$, $\hat{b} = \hat{h}_{\hat{\mu}}\{2(n\hat{h}_{\hat{\mu}}^3)^{-1} k_1 + 3\hat{h}_{\hat{\mu}}^2 k_2^2 \hat{J}_2\}$ (an estimator of $h_\mu M''(h_\mu)$), $W(u) = K(u) + uK'(u)$,

$$\hat{a}(x|h) = (nh)^{-1} \sum_{i=1}^{n} W\{(x - X_i)/h\}\,,$$

$$\hat{h}_{\hat{\nu}} = \hat{h}_{\hat{\mu}} + \hat{b}^{-1} \left\{ 2 \int \hat{f}(x|\hat{h}_{\hat{\mu}})\, \hat{a}(x|\hat{h}_{\hat{\mu}})\, dx - 2k_2\, \hat{h}_{\hat{\mu}}^2\, \hat{J}_1 \right\}. \tag{2.5}$$

Then $\hat{h} = \hat{h}_{\hat{\nu}}$ asymptotically minimizes the variance of $\hat{h} - \hat{h}_\nu$, over all possible data-based choices of $\hat{h}$.

3. Theoretical analysis

3.1 *Summary.* In subsection 3.2 we provide a simple, heuristic argument which introduces our main results and their implications. That argument involves positive constants σ^2, which are proportional to the asymptotic variances of quantities such as $\hat{h} - \hat{h}_\nu$ and $\hat{h} - h_\mu$. Those quantities are listed in subsection 3.3.

Unfortunately, the heuristic argument in subsection 3.2 is invalid in at least two fundamental respects. These difficulties are discussed in subsection 3.4, where we point out that in many instances of practical interest, either $E(\hat{h}^{-1}) = \infty$ or $E(\hat{h}^2) = \infty$. The first of these problems can result in both $\mu(\hat{h})$ and $\nu(\hat{h})$ being infinite. Its impact may be reduced by truncating the value of $\hat{h}$ when this quantity takes values too close to zero. Difficulties created by infiniteness of $E(\hat{h}^2)$ do not require truncation, but do demand considerable care in the formulation of our main results. For example, the value of $E(\hat{h} - \hat{h}_\nu)^2$ in a Taylor expansion from subsection 3.2 should be replaced by the asymptotic variance of $\hat{h} - \hat{h}_\nu$, which can be finite even when the former quantity is infinite. A rigorous formulation of our main results along these lines is given in subsection 3.5. Proofs are deferred to Section 5.

3.2 *Heuristic argument.* The discussion in this section is based on simple Taylor expansion, and describes the main features of $\mu(\hat{h})$ and $\nu(\hat{h})$, where $\hat{h}$ denotes a general data-driven bandwidth. Since $M'(h_\mu) = 0$ then

$$\begin{aligned}\mu(\hat{h}) &= E\{M(\hat{h})\} = E\{M(h_\mu) + (\hat{h} - h_\mu)\, M'(h_\mu) + \tfrac{1}{2}(\hat{h} - h_\mu)^2\, M''(h_\mu) + \ldots\} \\ &= M(h_\mu) + \tfrac{1}{2}\, E(\hat{h} - h_\mu)^2\, M''(h_\mu) + \ldots . \end{aligned} \tag{3.1}$$

Likewise, since $\Delta'(\hat{h}_\nu) = 0$ then

$$\begin{aligned}\nu(\hat{h}) &= E\{\Delta(\hat{h})\} = E\{\Delta(\hat{h}_\nu) + (\hat{h} - \hat{h}_\nu)\, \Delta'(\hat{h}_\nu) + \tfrac{1}{2}(\hat{h} - \hat{h}_\nu)^2\, \Delta''(\hat{h}_\nu) + \ldots\} \\ &= E\{\Delta(\hat{h}_\nu)\} + \tfrac{1}{2}\, E(\hat{h} - \hat{h}_\nu)^2\, M''(h_\mu) + \ldots . \end{aligned} \tag{3.2}$$

(To appreciate the last identity, note that $\{\Delta''(\hat{h}_\nu) - M''(h_\mu)\}/M''(h_\mu) \to 0$ in probability as $n \to \infty$. In both (3.1) and (3.2), the terms represented by "..." are of smaller order than the last included terms.) Replacing $\hat{h}$ by h_μ in (3.2) we deduce that

$$M(h_\mu) = E\{\Delta(\hat{h}_\nu)\} + \tfrac{1}{2}\, E(\hat{h}_\nu - h_\mu)^2\, M''(h_\mu) + \ldots .$$

Therefore, (3.2) entails

$$\nu(\hat{h}) = M(h_\mu) + \tfrac{1}{2}\,\{E(\hat{h} - \hat{h}_\nu)^2 - E(\hat{h}_\nu - h_\mu)^2\}\, M''(h_\mu) + \ldots . \tag{3.3}$$

Formulae (3.1) and (3.3) represent our main results. As we shall note in subsection 3.4, they must be modified if they are to hold rigorously, but their main features are preserved by those adjustments.

We may deduce from (3.1) that if $\hat{h}$ denotes a practical, data-driven bandwidth then $\mu(\hat{h})$ always exceeds $M(h_\mu)$. Of course, this is a trivial consequence of the definition of h_μ, but it does contrast with (3.3), which indicates that there is at least a potential for $\nu(\hat{h})$ to be less than $M(h_\mu)$. We shall show shortly that this potential can be realized in practice.

To more fully appreciate the implications of (3.1) and (3.3), let $\hat{h}_I$ and $\hat{h}_{II}$ denote two different, competing versions of $\hat{h}$. Then by (3.1),

$$\mu(\hat{h}_I) - \mu(\hat{h}_{II}) \sim \tfrac{1}{2}\,\{E(\hat{h}_I - h_\mu)^2 - E(\hat{h}_{II} - h_\mu)^2\}\, M''(h_\mu)\,, \tag{3.4}$$

while by (3.3),

$$\nu(\hat{h}_I) - \nu(\hat{h}_{II}) \sim \tfrac{1}{2}\,\{E(\hat{h}_I - \hat{h}_\nu)^2 - E(\hat{h}_{II} - \hat{h}_\nu)^2\}\, M''(h_\mu)\,. \tag{3.5}$$

We claim that in cases of practical interest, it is possible for the right-hand sides of (3.4) and (3.5) to be of opposite signs, indicating that as "performance measures", $\mu(\hat{h})$ and $\nu(\hat{h})$ can provide quite different views of the suitability of different bandwidths.

For example, let $\hat{h}_I = \hat{h}_{\hat{\nu}}$ denote a bandwidth that is asymptotically optimal for approximating $\hat{h}_\nu$, and let $\hat{h}_{II} = \hat{h}_{pi,2}$ denote a plug-in bandwidth selector that has a relative error of $n^{-\frac{1}{2}}$ in terms of approximating h_μ. Then $\hat{h}_I - h_\mu$, $\hat{h}_{II} - h_\mu$, $\hat{h}_I - \hat{h}_\nu$, $\hat{h}_{II} - \hat{h}_\nu$ are of sizes $n^{-\frac{3}{10}}$, $n^{-\frac{1}{2}}$, $n^{-\frac{3}{10}}$ and $n^{-\frac{3}{10}}$, respectively. Writing $\sigma^2_{\mu,\nu}$, $\sigma^2_{\hat{\nu}}$, $\sigma^2_{\hat{\nu},\mu}$ for the limits of $n^{\frac{8}{5}} E(\hat{h}_\nu - h_\mu)^2$, $n^{\frac{8}{5}} E(\hat{h}_{\hat{\nu}} - \hat{h}_\nu)^2$, $n^{\frac{8}{5}} E(\hat{h}_{\hat{\nu}} - h_\mu)^2$ respectively, and letting c_3 denote the constant such that $M''(h_\mu) \sim c_3\, n^{-\frac{2}{5}}$, we deduce from (3.4) and (3.5) that

$$\mu(\hat{h}_I) - \mu(\hat{h}_{II}) \sim \tfrac{1}{2}\, \sigma^2_{\hat{\nu},\mu}\, c_3\, n^{-1}\,,$$
$$\nu(\hat{h}_I) - \nu(\hat{h}_{II}) \sim \tfrac{1}{2}\, (\sigma^2_{\hat{\nu}} - \sigma^2_{\mu,\nu})\, c_3\, n^{-1}\,.$$

Of course, $\sigma^2_{\hat{\nu},\mu} > 0$. We shall show in subsection 3.3 that $\sigma^2_{\hat{\nu}} < \sigma^2_{\mu,\nu}$. Therefore, $\mu(\hat{h}_I) < \mu(\hat{h}_{II})$ and $\nu(\hat{h}_I) > \nu(\hat{h}_{II})$, for large n. That is, the criteria μ and ν provide diametrically opposite views of the relative performance of $\hat{h}_I$ and $\hat{h}_{II}$.

Note too that when $\hat{h} = \hat{h}_{\hat{\nu}}$ we have by (3.3),

$$\nu(\hat{h}) - M(h_\mu) \sim \tfrac{1}{2}\, (\sigma^2_{\hat{\nu}} - \sigma^2_{\mu,\nu})\, c_3\, n^{-1} < 0\,.$$

Hence, for large n, $\nu(\hat{h}) < M(h_\mu)$, indicating that practical, data-driven bandwidths can produce an expected integrated squared error that is strictly less than the minimum of mean integrated squared error. When one appreciates that $\hat{h}_{\hat{\nu}}$ is a data-driven attempt at minimizing Δ, not M, this result is not counter-intuitive. Nevertheless, our discussion of the result with colleagues indicates that many find it surprising.

3.3 *Values of* σ^2. As indicated in Section 2, we study five different data-driven bandwidths, $\hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$. Of these, all but $\hat{h}_\nu$ represent practical choices of h. It may be proved that the quantities

$$n^{\frac{3}{10}}(\hat{h}_\nu - h_\mu)\,, \quad n^{\frac{3}{10}}(\hat{h}_{\hat{\nu}} - \hat{h}_\nu)\,, \quad n^{\frac{3}{10}}(\hat{h}_{cv} - \hat{h}_\nu)\,,$$
$$n^{\frac{3}{10}}(\hat{h}_{\hat{\nu}} - h_\mu)\,, \quad n^{\frac{3}{10}}(\hat{h}_{cv} - h_\mu)\,, \quad n^{\frac{7}{10}}(\hat{h}_{pi,2} - h_\mu)$$

are asymptotically normally distributed with positive variances $\sigma^2_{\mu,\nu}$, $\sigma^2_{\hat\nu}$, $\sigma^2_{cv,\nu}$, $\sigma^2_{\hat\nu,\mu}$, $\sigma^2_{cv,\mu}$, $\sigma^2_{pi,2}$ respectively. The asymptotic normality does not directly concern us here, but the values of σ^2 do, and so we list them below. They are calculated from results of Hall and Marron (1987a), Park and Marron (1990), Fan and Marron (1992), Hall *et al.* (1991), Hall and Marron (1991), Jones, Marron and Park (1991), Jones and Sheather (1991) and Hall and Johnstone (1992).

Let $J_j = \int (f^{(j)})^2$, $I_j = \int (f^{(j)})^2 f$, $k_1 = \int K^2$, $k_2 = \int u^2 K(u)\,du$, $k_3 = \int [\int K(u+v)\{K(v) + vK'(v)\}\,dv]^2\,du$, $k_5 = \int u^2 K'(u)^2\,du$,

$$k_6 = \int \{K(u) + uK'(u)\}\,du \int K(u+v)\{K(v) + vK'(v)\}\,dv\,.$$

Then

$$\begin{aligned}
\sigma^2_{\mu,\nu} &= \tfrac{8}{25}\left(k_1^{-7} k_2^{-6}\right)^{\frac{1}{5}} k_3\, J_0\, J_2^{-\frac{3}{5}} + \tfrac{4}{25}\left(k_1\, k_2^3\right)^{-\frac{2}{5}} J_2^{-\frac{8}{5}}\left(I_2 - J_1^2\right), \\
\sigma^2_{\hat\nu} &= \tfrac{4}{25}\left(k_1\, k_2^3\right)^{-\frac{2}{5}} J_2^{-\frac{8}{5}}\left(I_2 - J_1^2\right), \\
\sigma^2_{cv,\nu} &= \tfrac{8}{25}\left(k_1^{-7} k_2^{-6}\right)^{\frac{1}{5}} k_5\, J_0\, J_2^{-\frac{3}{5}} + \tfrac{4}{25}\left(k_1^{-1} k_2^{-3}\right)^{\frac{2}{5}} J_2^{-\frac{8}{5}}\left(I_2 - J_1^2\right), \\
\sigma^2_{\hat\nu,\mu} &= \tfrac{8}{25}\left(k_1^{-7} k_2^{-6}\right)^{\frac{1}{5}} k_3\, J_0\, J_2^{-\frac{3}{5}}, \\
\sigma^2_{cv,\mu} &= \sigma^2_{\mu,\nu} + \sigma^2_{cv,\nu} - 2\tau, \\
\sigma^2_{pi,2} &= \tfrac{4}{25}\left(k_1\, k_2^{-2}\right)^{\frac{2}{5}} J_2^{-\frac{12}{5}}\left(I_4 - J_2^2\right),
\end{aligned}$$

where τ has the same definition as $\sigma^2_{\mu,\nu}$ except that k_6 replaces k_3.

The situation is a little different in the case of the plug-in bandwidth rule $\hat h_{pi,1}$. There, $n^{\frac{8}{5}}\,(\hat h_{pi,1} - h_\mu)^2$ converges in probability to $\sigma^2_{pi,1} > 0$; the asymptotic distribution, at this level, is not normal $N(0,1)$. Defining $k_4 = \int u^4 K(u)\,du$, the value of σ^2 is given by

$$\sigma^2_{pi,1} = \tfrac{1}{400}\left(k_1^3\, k_2^{-11}\right)^{\frac{2}{5}} k_4^2\, J_2^{-\frac{16}{5}}\, J_3^2\,;$$

see Hall *et al.* (1991).

Note particularly that $\sigma^2_{\hat\nu} < \sigma^2_{\mu,\nu}$, as argued in subsection 3.2.

3.4 *Difficulties with the heuristic argument.* The argument given in subsection 3.2 is of course non-rigorous. It is incorrect in at least two important respects. Firstly, the condition

$$E(\hat{h}^{-1}) < \infty \tag{3.6}$$

is necessary and sufficient for even the finiteness of $\mu(\hat{h})$ and $\nu(\hat{h})$, let alone for the validity of Taylor expansions such as those derived in subsection 3.2. The condition can be violated by practical bandwidth selection rules. Secondly, Taylor expansions such as (3.1)–(3.5) require at least the finiteness of $E(\hat{h}^2)$, and there is no guarantee that this condition holds. Indeed, we shall prove shortly that for the cross-validation bandwidth $\hat{h}_{cv}$, it is possible to choose a symmetric nonnegative kernel function K that is bounded, smooth and compactly supported, yet has the property

$$P(\hat{h}_{cv} = \infty) > 0 \quad \text{for all } n \geq 2 \text{ and all densities } f \text{ that are supported on the entire real line.} \tag{3.7}$$

Perhaps unexpectedly, this pessimistic result does not significantly alter our main conclusions such as (3.2) and (3.3). To appreciate why, note that a condition such as $P(\hat{h} = \infty) > 0$ does not affect the finiteness of either $\mu(\hat{h})$ or $\nu(\hat{h})$. However, it is clear that at the very least, our notation in results such as (3.2) and (3.3) must change. In particular, it is not permissible to write down terms such as $E(\hat{h} - h_\mu)^2$, which may be infinite. Instead we must work with the asymptotic variances σ^2, described in subsection 3.3. This has important implications for our proofs, which are significantly more complex than might be expected from the simple discussion in subsection 3.2.

Let us return to the problems caused by infiniteness of $E(\hat{h}^{-1})$. Simple calculations show that for all $h > 0$,

$$\tfrac{1}{2}(nh)^{-1} \int K^2 - \int f^2 \leq \Delta(h) \leq 2h^{-1} \int K^2 + 2 \int f^2 , \tag{3.8}$$

from which it follows that condition (3.6) is necessary and sufficient for $\mu(\hat{h}) < \infty$ and for $\nu(\hat{h}) < \infty$. It may be shown that (3.6) holds in the case $\hat{h} = \hat{h}_{cv}$, although this is not a trivial matter; see Lemma 5.1 in Section 5. Since $\Delta(\hat{h}_\nu) \le \Delta(\hat{h}_{cv})$ then $E(\hat{h}_{cv}^{-1}) < \infty$ implies $E(\hat{h}_\nu^{-1}) < \infty$. However, in the cases $\hat{h} = \hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$, finiteness of $E(\hat{h}^{-1})$ is at least partly a matter of definition. The definitions given in Section 2 do not even guarantee that $\hat{h}$ is real-valued and positive, and a density estimator with other than a positive bandwidth is not really properly defined. Admittedly, the probability that $\hat{h}$ is not positive is exponentially small, but since it can be nonzero in each of the cases $\hat{h} = \hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$ then practical problems arise in defining both $\mu(\hat{h})$ and $\nu(\hat{h})$. Obviously, simply replacing $\hat{h}$ by zero in the event that $\hat{h}$ is not positive results in failure of condition (3.6). An alternative is to replace $\hat{h}$ by

$$\tilde{h} = \begin{cases} \hat{h} & \hat{h} > n^{-r} \\ n^{-r} & \text{otherwise}\,, \end{cases} \tag{3.9}$$

for any $r > \frac{1}{5}$. This guarantees (3.6). The modification (3.9) is appropriate for all our purposes, regardless of how large the value of r might be. In practice, it amounts to modifying $\hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$ so that those quantities are no more than algebraically small. In the case of $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$, it is sufficient to modify the definitions at (2.3) and (2.4) to

$$\hat{h}_{pi,1} = \kappa_1(n|\hat{J}_2|)^{-\frac{1}{5}}\,, \qquad \hat{h}_{pi,2} = \kappa_1(n|\hat{J}_2|)^{-\frac{1}{5}} + \kappa_2\,|\hat{J}_2|^{-\frac{8}{5}}\,|\hat{J}_3|^{-\frac{3}{5}}\,. \tag{3.10}$$

Finally in this subsection, we verify condition (3.7). The kernel which we produce has $K(0) = 0$, but less extreme examples with moments of $\hat{h}_{cv}$ infinite are possible. We shall check (3.7) only for the range $n \ge n_0$, where $n_0 \ge 1$ is sufficiently large and does not depend on f. The case $1 \le n \le n_0$ may be handled by a simple subsidiary argument.

Let the symmetric probability density K vanish in a neighborhood of the origin and outside the interval $(-1, 1)$, and have the property

$$\int K^2 - 2\int |y|^{-1}\,K(y)\,dy > 0\,. \tag{3.11}$$

This may be achieved by giving K two short, sharp peaks on either side of the origin. Of course, K may be as smooth as we like, in the sense of having an arbitrary number of bounded derivatives. Since

$$\int \hat{f}(x)^2\,dx \geq (nh)^{-2} \sum_{i=1}^{n} \int K\{(x - X_i)/h\}^2\,dx = (nh)^{-1} \int K^2\,,$$

then

$$nh\,\mathrm{CV}(h) \geq \int K^2 - 4(n-1)^{-1}\,S_n(h) \tag{3.12}$$

where

$$S_n(h) = \sum_{1 \leq i < j \leq n} \sum K\{(X_i - X_j)/h\}\,.$$

Let $X_j \in \mathcal{I}_j = (e^j - 1, e^j + 1)$, $j \geq 1$. Then $X_j - X_i = e^j(1 - e^{i-j}) + U_{ij}$ where $|U_{ij}| \leq 2$. In this notation,

$$S_n(h) = \sum_{1 \leq i < j \leq n} \sum K[\{e^j(1 - e^{i-j}) + U_{ij}\}\,h^{-1}]\,.$$

We shall prove by contradiction that for some $n_0 \geq 2$, we have for all $h_\mu > 0$,

$$4(n-1)^{-1} \sum_{|U_{ij}| \leq 2} \sup_{0 < h \leq h_\mu} S_n(h) < \int K^2\,, \qquad \text{all } n \geq n_0\,. \tag{3.13}$$

If this result were false then we could choose $h_n \to \ell$, $0 \leq \ell \leq \infty$, and numbers $|U_{ij}| \leq 2$, such that along an infinite sequence of n's we have $n^{-1} \log h_n \to t$, $-\infty \leq t \leq \infty$, and

$$4(n-1)^{-1}\,S_n(h_n) \geq \int K^2\,. \tag{3.14}$$

To simplify notation we assume that the subsequence is the whole sequence; the proof for a proper subsequence is identical. If $\ell < \infty$ then $S_n(h_n)$ is bounded in n, and so (3.14) must fail. Therefore, $\ell = \infty$ and $0 \leq t \leq \infty$.

If $t > 1$ then $e^n/h_n \to 0$, whence

$$\sup_{1 \leq i < j \leq n} \{e^j(1 - e^{i-j}) + |U_{ij}|\}/h_n \to 0\,.$$

Therefore, since K vanishes in a neighborhood of the origin, $S_n(h_n) = 0$ for all sufficiently large n. Again, (3.14) fails. Hence $\ell = \infty$ and $0 \le t \le 1$.

When $t \le 1$ we argue that

$$\begin{aligned} S_n(h_n) &\le \sum\sum_{1 \le i < j < \infty} K[\{e^j(1 - e^{i-j}) + U_{ij}\}\ {h_n}^{-1}] \\ &\sim \sum_{j=1}^{\infty} jK(e^j/h_n) \\ &\sim \int_0^{\infty} xK(e^x/h_n)\,dx \sim (\log h_n) \int_0^{\infty} y^{-1}\,K(y)\,dy\,. \end{aligned}$$

Consequently,

$$\limsup_{n \to \infty}\ n^{-1}\,S_n(h_n) \le \int_0^{\infty} y^{-1}\,K(y)\,dy = \tfrac{1}{2} \int |y|^{-1}\,K(y)\,dy\,.$$

In view of (3.14) this entails

$$2 \int |y|^{-1}\,K(y)\,dy \ge \int K^2\,,$$

which contradicts (3.11). Therefore, (3.13) must hold.

Noting (3.12), we see that (3.13) implies

$$P\{\mathrm{CV}(h) > 0\,,\ \text{all } h > 0\} \ge P(X_j \in \mathcal{I}_j,\ 1 \le j \le n)\,,\ \text{all } n \ge n_0\,.$$

Since $\mathrm{CV}(\infty) = 0$ then

$$P(\hat{h}_{cv} = \infty) \ge \prod_{j=1}^{n} P(X_j \in \mathcal{I}_j) > 0$$

if $f > 0$, as had to be shown.

3.5 *Rigorous formulation.* Here we establish rigorous versions of our principal expansions (3.1) and (3.3), leading directly to rigorous versions of (3.4) and (3.5). As shown in subsection 3.4, we must replace mean squared errors in those formulae by their asymptotic counterparts. We do this for all the empirical bandwidths introduced in Section 2.

Throughout it is assumed that

K is a symmetric probability density, vanishing outside $(-1,1)$,

having two derivatives on $(-\infty,\infty)$ and satisfying

$$|K''(x+y)-K''(x)| \le C|y|^{\eta} \text{ for some } C,\eta>0 \text{ and all } x,y\,; \tag{3.15}$$

and

f and its first two derivatives are bounded and integrable on $(-\infty,\infty)$,

and f'' satisfies $|f''(x+y)-f''(x)| \le C|y|^{\eta}(1+|x|)^{-c}$ (3.16)

for some $c>\frac{1}{2}$, $C,\eta>0$ and all x,y.

Additional conditions on f are needed to ensure that specific versions of $\hat{h}$ enjoy the properties ascribed to them in Section 2. In the case $\hat{h}=\hat{h}_{pi,1}$ or $\hat{h}=\hat{h}_{pi,2}$, we ask that

f has four bounded, integrable derivatives on $(-\infty,\infty)$, and

$$|f^{(4)}(x+y)-f^{(4)}(x)| \le C|y|^{\eta} \tag{3.17}$$

for some $C>0$ and $\frac{1}{4}<\eta\le 1$, and all x,y.

When $\hat{h}=\hat{h}_{\hat{\nu}}$ we ask that (3.17) hold but that the range of values of η be reduced to $\frac{1}{2}<\eta\le 1$.

For the cases $\hat{h}=\hat{h}_{\nu}$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$, let $\alpha=\frac{3}{10},\frac{3}{10},\frac{3}{10},\frac{3}{10},\frac{3}{5}$ and $\frac{1}{2}$, respectively. (Then, $\hat{h}-h_{\mu}$ is of size $n^{-\alpha}$.) In the theorem below we assume (3.15) and (3.16) throughout. When $\hat{h}=\hat{h}_{pi,1}$ or $\hat{h}_{pi,2}$, we assume too that (3.17) holds, and when $\hat{h}=\hat{h}_{\hat{\nu}}$, we assume the version of (3.17) in which $\frac{1}{2}<\eta\le 1$. In the case $\hat{h}=\hat{h}_{\hat{\nu}}$, the definition of $\hat{h}_{\hat{\nu}}$ given in Section 2 should be modified as indicated at (3.9), for the reasons given there. (In that formula, any $r>\frac{1}{5}$ may be used.) When $\hat{h}=\hat{h}_{pi,1}$ or $\hat{h}_{pi,2}$, either of the modifications in (3.9) or (3.10) may be employed. Recall that $M''(h_{\mu})\sim c_3\,n^{-\frac{2}{5}}$, where $c_3>0$ is given in Section 2.

The following result presents rigorous versions of (3.1) and (3.3).

THEOREM. *Assume the conditions prescribed above. Let* $\hat{h} = \hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi,1}$ *or* $\hat{h}_{pi,2}$.

(a) *Take* $\sigma^2 = \sigma^2_{\mu,\nu}$, $\sigma^2_{\hat{\nu},\mu}$, $\sigma^2_{cv,\mu}$, $\sigma^2_{pi,1}$ *or* $\sigma^2_{pi,2}$, *respective to the values of* $\hat{h}$ *listed above. Then*

$$\mu(\hat{h}) = M(h_\mu) + \tfrac{1}{2}\,\sigma^2\,c_3\,n^{-2\alpha-\frac{2}{5}} + o(n^{-2\alpha-\frac{2}{5}}) \tag{3.18}$$

as $n \to \infty$.

(b) *Take* $\sigma^2 = 0$, $\sigma^2_{\hat{\nu}}$, $\sigma^2_{cv,\nu}$, $\sigma^2_{\mu,\nu}$ *or* $\sigma^2_{\mu,\nu}$, *respective to the values of* $\hat{h}$ *listed above. Then*

$$\nu(\hat{h}) = M(h_\mu) + \tfrac{1}{2}\,(\sigma^2 - \sigma^2_{\mu,\nu})\,c_3\,n^{-1} + o(n^{-1}) \tag{3.19}$$

as $n \to \infty$.

It is straightforward to use (3.18) and (3.19) to derive rigorous versions of the results discussed formally in subsection 3.2. For example, take $\hat{h}_I = \hat{h}_{\hat{\nu}}$ and $\hat{h}_{II} = \hat{h}_{pi,2}$. Then by (3.18) and (3.19),

$$\mu(\hat{h}_I) - \mu(\hat{h}_{II}) \sim \tfrac{1}{2}\,\sigma^2_{\hat{\nu},\mu}\,c_3\,n^{-1} > 0\,,$$
$$\nu(\hat{h}_I) - \nu(\hat{h}_{II}) \sim \tfrac{1}{2}\,(\sigma^2_{\hat{\nu}} - \sigma^2_{\mu,\nu})\,c_3\,n^{-1} < 0\,,$$

precisely as suggested in subsection 3.2. Arguing in this manner, the conclusions drawn there may be rigorously established.

Both (3.18) and (3.19) may be carried to higher-order terms, using methods of proof similar to those given in Section 5. However, the results in our present theorem are adequate to confirm the validity of the conclusions drawn earlier by heuristic argument.

4. Simulations

The theoretical results concerning the different ranking of bandwidth selectors with respect to the risks μ and ν are based on refined bandwidth estimators, designed to approach either $\hat{h}_\nu$ or h_μ with fast convergence rates. A major problem

in the practical implementation of such bandwidth procedures, as for example $\hat{h}_{\hat{\nu}}$ or $\hat{h}_{\hat{\mu}}$, is that they are based on estimates of J_1, J_2, and/or J_3. These quantities are hard to estimate, see Aldershof (1991), Hall and Marron (1987b), Hall *et al.* (1991) and Jones, Marron and Park (1991), which might cause serious problems under small samples. Moreover, the theoretical comparison of the μ and ν risks involves only the leading terms of their asymptotic representations. In order to get an impression on the relative risk behaviour for small and moderate sample sizes, we have performed some simulations. We have addressed the following problems.

1. Is there a relevant practical difference in assessing risk via μ or ν for small to moderate sample sizes? If so, under which conditions?

2. Do either μ or ν tailored bandwidth selectors have an advantage for small to moderate sample sizes? When the behaviour is about the same, do ν-oriented bandwidth procedures get more or less competitive with increasing sample size?

3. The bandwidth estimators cited here are constructed using the leading terms of the asymptotic representation of either μ or ν, rather than the actual risks. To which extent do the leading terms reflect the actual μ and ν risks? In case there are significant differences, how well do the risks of the resulting kernel estimators approach the optimal actual risks, as opposed to the optimal asymptotic risks?

We implemented several bandwidth selectors, including a least-squares cross-validation bandwidth and several plug-in rules, and investigated their μ and ν behavior on a variety of densities for samples of size $n = 100$ and $n = 1000$. In order to ensure a wide range of underlying densities, we used the collection of 15 mixtures of normal distributions, suggested by Marron and Wand (1992).

As estimators targeting on ν-optimality we used several versions of a bandwidth selector proposed by Hall and Johnstone (1992), in the present paper given by formula (2.5). The bandwidth $\hat{h}_{HJS1}$ is a truly data-driven adaptation of (2.5).

Here, we used for the pilot bandwidth $\hat{h}_{\hat{\mu}}$ essentially the plug-in estimator proposed by Sheather and Jones (1992), although we based "scale" on the sample standard deviation instead on the interquartile range. This version of the pilot does not have root-n asymptotics, but in the simulations of Jones, Marron and Sheather (1992) it performed typically superior to the root-n methods, for the sample sizes considered here. The quantities J_1 and J_2 are estimated by one-stage estimators following Jones and Sheather (1991). This involves oversmoothed kernel estimators for integrated squared derivatives of f, with the bandwidth chosen using a normal reference distribution. For details see Park and Marron (1992) and Aldershof (1991).

Among the procedures aiming at μ-optimality we focussed on the bandwidth $\hat{h}_{PMPI}$, a plug-in estimator introduced by Park and Marron (1990), as previous simulations have shown this to be near the middle (hence a reasonable representative) of proposed methods, in the V_μ context.

In Figure 1 we plot the empirical μ and ν risks of the bandwidth selectors $\hat{h}_{PMPI}$ and $\hat{h}_{HJS1}$, constructed to minimize μ and ν, respectively, for the densities #1 - #15 in Marron and Wand (1992). The empirical risks are computed by averaging over the (exact) M and Δ values obtained for 500 samples of size $n = 100$ and $n = 1000$ in Figures 1a and 1b, respectively. Note that the underlying densities are listed on the horizontal axis. In Figure 1a, $n = 100$, the densities are ordered with respect to increasing $\nu(\hat{h}_\nu)$, the optimal ν value; i. e. from the left to the right it gets more and more difficult to estimate a particular density by a kernel estimator. This ordering is maintained later to ensure comparability.

The most striking feature in Figure 1 is that we have two pairs of almost coinciding curves: the empirical μ and ν values of $\hat{h}_{PMPI}$, represented by the dash-and-dotted line and the solid line, are very similar, and so are the empirical μ and ν values of $\hat{h}_{HJS1}$, given by the dotted lines. This indicates that for $n = 100$, or even $n = 1000$, it is not really important whether we assess the performance of

bandwidth selectors by ν or by μ, despite the major philosophical differences, and contrary to the asymptotical theory. In the case $n = 100$, for 9 of the 15 densities the empirical μ and ν risks of $\hat{h}_{PMPI}$ and $\hat{h}_{HJS1}$ are all very close, so that here even the choice of the bandwidth procedure has not much influence on the final behaviour of the kernel estimator. This is true for densities that are "easy" to estimate, as the normal density #1, and the bimodal and trimodal densities #6, #8 and #9, as well as the double claw densities #11 and #13, but the more tricky claw densities #10 and #12 are included as well. For $n = 1000$, however, there is a substantial difference, with $\hat{h}_{HJS1}$ being significantly worse in both senses. But the important point is that it makes no difference whether the information contained in the data is filtered through the loss M or through loss Δ. In both cases the data seem to reflect the same features of the underlying distribution.

The results of the simulation study are somewhat surprising with respect to the ranking of the bandwidth selectors, at least at the first view. We know that for large samples the bandwidth $\hat{h}_{HJS1}$ has a smaller ν-risk than $\hat{h}_{PMPI}$, whereas μ prefers $\hat{h}_{PMPI}$. This follows directly from the Theorem in Section 3, one of the main results of the present paper. In our study, however, μ and ν provide essentially the same ranking of both methods, for all densities and for both $n = 100$ and $n = 1000$. That means that for comparison of $\hat{h}_{HJS1}$ and $\hat{h}_{PMPI}$ the asymptotic lessons are not dominant, even for the (often considered) rather large sample size of $n = 1000$.

However, the poor behavior of $\hat{h}_{HJS1}$, even in the V_ν sense, relative to $\hat{h}_{PMPI}$ does not necessarily imply the V_ν approach is hopeless, for samples of these sizes. We believe that the problems with $\hat{h}_{HJS1}$ are mostly caused by the altogether poor performance of kernel estimators for J_1 and J_2, as shown in Aldershof (1991), and up to a certain degree by the choice of the pilot bandwidth $\hat{h}_{\hat{\mu}}$. Hence we tried various combinations of using the true values of J_1 and J_2, and also h_μ to develop "theoretical versions" of $\hat{h}_{\hat{\nu}}$. Generally, we observed that using the theoretical $\hat{h}_\nu$ gave only marginal improvement, but the theoretical values of J_1 and J_2 gave

considerably smaller risks. To save space, we omit the details here.

Now we compare $\hat{h}_{PMPI}$ with the best of these theoretical versions of $\hat{h}_\nu$, bandwidth $\hat{h}_{HJTT}$, which uses both the pilot bandwidth h_μ, and also the true J_1 and J_2. Figures 2a and 2b give the results for $n = 100$ and $n = 1000$, respectively. We find that the ν-tailored bandwidth $\hat{h}_{HJTT}$ is generally superior, in both senses, to the μ-oriented $\hat{h}_{PMPI}$. For sample size $n = 100$ the procedure $\hat{h}_{HJTT}$ is uniformly better, in the ν-sense, than $\hat{h}_{PMPI}$ over the whole catalog of densities, with more improvements for the "difficult" densities #3, #4, #10, #12, #14 and #15. The picture changes only slightly for the higher sample size $n = 1000$. The only major difference to $n = 100$ is that $\hat{h}_{PMPI}$ is now competing with $\hat{h}_{HJTT}$ for the kurtotic unimodal density #4, too. Note that for the easier to estimate densities #1, #2 and #6, there is a tendency for the theoretically predicted result, that $\hat{h}_{HJTT}$ would be better with respect to ν, but $\hat{h}_{PMPI}$ better with respect to μ, to hold. This tendency for "asymptotics to kick in earlier for the easier densities" was also observed in the simulations of Jones, Marron and Sheather (1992).

We included in all figures $\nu(\hat{h}_\nu)$, the true values of the optimal ν-risk, as well as the values of the leading term of $\nu(\hat{h}_\nu)$, or the asymptotic optimal risk. This gives an indication of when (i.e. for which densities) the asymptotic lessons are far from holding for these sample sizes. The asymptotic optimum is especially far from the true optimum for those densities with strong spikes. This is because the asymptotic optimum is based on integrated squared density derivatives, which are very large in these cases, but not important components of the true optimum.

It is interesting to see that the empirical risks of both versions $\hat{h}_{HJTT}$ and $\hat{h}_{PMPI}$ are much closer to the *true* optimal risk ν than to the *asymptotic* optimal risks. We might have expected otherwise, as the bandwidth procedure $\hat{h}_{HJTT}$ has been derived to minimize the leading terms of the Taylor expansion of $\nu(\hat{h}_\nu)$, so that it is actually targeted on the asymptotic optimal risk. This result might

encourage us to develop bandwidth procedures by minimizing estimated asymptotic risks. Even in cases where the leading terms of the Taylor expansion differ strongly from the true risk, the bandwidth selector may still behave well in terms of the empirical risk.

Summary. The simulation study gave us the impression that for small or moderate samples it is of minor importance whether the behaviour of bandwidth selectors is measured by the risk μ or by the risk ν, as both values appeared to be close in all cases.

It is a matter of personal taste whether a statistician advocates V_μ or V_ν. The viewpoints reflect different decision theoretic problems, and none of the two is a priori superior to the other. However, the different philosophical approaches in V_μ and V_ν suggest the use of corresponding data-driven bandwidth procedures, and in spite of asymptotic optimality we see for practical relevant sample sizes the ν-tailored bandwidth $\hat{h}_{HJS1}$ in strong disadvantage compared to the μ-motivated $\hat{h}_{PMPI}$. In general, we have the impression that *in the current state of art* bandwidth procedures developed under V_μ are the better choice for practical problems, independent of our philosophical point of view. Nevertheless, there is much space to improve on ν-oriented, honestly data-driven bandwidth selectors, so that in a short time there may be more practical motivation to ask again the question: "μ or ν"?

5. Proof of Theorem. We derive only (3.19), as a proof of (3.18) is similar but simpler. Define $\alpha = \frac{3}{10}$ in the cases $\hat{h} = \hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$ or $\hat{h}_{cv}$, put $\alpha = \frac{3}{5}$ when $\hat{h} = \hat{h}_{pi,1}$, and let $\alpha = \frac{1}{2}$ when $\hat{h} = \hat{h}_{pi,2}$. Put $a = \alpha - \epsilon$, where $\epsilon \in \left(0, \frac{1}{100}\right)$ is fixed but arbitrarily small. Let $b \gg a$ and $C > 0$ be fixed positive constants. Our proof of the theorem involves dividing the range of values $\hat{h}$ can take into four parts: $\hat{h} \le n^{-b}$, $\hat{h} > C$, $n^{-b} < \hat{h} \le C$ but $|\hat{h} - h_\mu| > n^{-a}$, and $|\hat{h} - h_\mu| \le n^{-a}$,respectively. The first two lemmas below describe properties of $\Delta(\hat{h})$ when $\hat{h}$ is confined to the first two of these ranges, and when $\hat{h} = \hat{h}_{cv}$. The third lemma shows that

$P(|\hat{h}_{cv} - h_\mu| > n^{-a}) = O(n^{-\lambda})$ for all $\lambda > 0$, and the fourth lemma states an identical result in the case where $\hat{h}_{cv}$ is replaced by any one of $\hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$, $\hat{h}_{pi,2}$. (The third and fourth lemmas are used to describe properties of $\Delta(\hat{h})$ when $\hat{h}$ lies in the third of the four ranges described above.) The fifth lemma provides basic properties of the stochastic processes Δ, Δ' and Δ'', and the sixth lemma gives formulae for $E(\hat{h} - h_\mu)^2$ and $E(\hat{h} - \hat{h}_\nu)^2$. The proof of the theorem is completed following proofs of the lemmas.

Recall that

$$\mathrm{CV}(h) = \int \hat{f}_h(x)^2\,dx - 2\{n(n-1)h\}^{-1} \sum\sum_{i \neq j} K\{(X_i - X_j)/h\}$$

is the cross-validation criterion.

LEMMA 5.1. *Let $\hat{h}_{cv}$ denote a bandwidth that minimizes* CV. *Assume that K is bounded with support contained within $[-1, 1]$, and that f is bounded. Then for each $b > 0$,*

$$E\Big(\Big[\int \{\hat{f}(x|\hat{h}_{cv}) - f(x)\}^2\,dx\Big]\, I(\hat{h}_{cv} \leq n^{-b})\Big) = O\{(1 + n^{(3-b)/2})\, P(\hat{h}_{cv} \leq 2n^{-b})^{\frac{1}{2}}\}\,. \tag{5.1}$$

PROOF. Let $\mathcal{E}$ denote the event that three or more pairs (X_i, X_j), with $1 \leq i < j \leq n$, satisfy $|X_i - X_j| \leq \hat{h}_{cv}$. Write $\tilde{\mathcal{E}}$ for the complement of $\mathcal{E}$, let q denote the left-hand side of (5.1), put

$$q(\mathcal{A}) = E\Big(\Big[\int \{\hat{f}(x|\hat{h}_{cv}) - f(x)\}^2\,dx\Big]\, I(\mathcal{A}, \hat{h}_{cv} \leq n^{-b})\Big)$$

for events $\mathcal{A}$, and let $q_1 = q(\mathcal{E})$, $q_2 = q(\tilde{\mathcal{E}})$. Then $q = q_1 + q_2$, and so it suffices to prove that the bound (5.1) applies to both q_1 and q_2.

First we treat q_1. Let $k_0 = k_0(n)$ denote the largest integer not exceeding $b \log_2 n$ and put

$$p_n(u) = P\{\text{3 or more pairs } (X_i, X_j), \quad \text{with } 1 \leq i < j \leq n, \text{ satisfy } |X_i - X_j| \leq u\}\,.$$

Let $X_{(1)} \le \ldots \le X_{(n)}$ denote the order statistics of the sample $X_1, \ldots, X_n$, and let $S_i = X_{(i+1)} - X_{(i)}$, $0 \le i \le n-1$, represent the spacings. Write $S_{(1)} \le \ldots \le S_{(n-1)}$ for the order statistics of the spacings. Noting that f is bounded, and that spacings are asymptotically independent and exponentially distributed, we may prove that for a constant $C_1 > 0$,

$$p_n(u) \le P(S_{(3)} \le u) \le C_1 (nu)^3 . \tag{5.2}$$

Therefore, noting that

$$\int (\hat{f} - f)^2 \le 2\Big(h^{-1} \int K^2 + \int f^2\Big) ,$$

we see that there exists a constant $C_2 > 0$ such that

$$\begin{aligned}
q_1 &\le C_2 \, E\{\hat{h}_{cv}^{-1} \, I(\mathcal{E}, \hat{h}_{cv} \le n^{-b})\} \\
&\le C_2 \sum_{k=k_0}^{\infty} 2^{k+1} \, P\big(\mathcal{E}, 2^{-k-1} < \hat{h}_{cv} \le 2^{-k}\big) \\
&\le C_2 \sum_{k=k_0}^{\infty} 2^{k+1} \, \{p_n(2^{-k}) \, P\big(2^{-k-1} < \hat{h}_{cv} \le 2^{-k}\big)\}^{\frac{1}{2}} \\
&\le C_2 \Big\{ \sum_{k=k_0}^{\infty} 2^{2k+2} \, p_n(2^{-k}) \Big\}^{\frac{1}{2}} \, P(\hat{h}_{cv} \le 2^{-k_0})^{\frac{1}{2}} \\
&\le C_1^{\frac{1}{2}} \, C_2 \, n^{\frac{3}{2}} \Big(\sum_{k=k_0}^{\infty} 2^{2k+2-3k} \Big)^{\frac{1}{2}} \, P(\hat{h}_{cv} \le 2^{-k_0})^{\frac{1}{2}} \\
&\le C_3 \, n^{(3-b)/2} \, P\big(\hat{h}_{cv} \le 2n^{-b}\big)^{\frac{1}{2}} .
\end{aligned}$$

Next we treat q_2. Suppose that at most two pairs (X_i, X_j), with $1 \le i < j \le n$, have $|X_i - X_j| \le \hat{h}_{cv}$. Then, for $n \ge 5$,

$$\{n(n-1)\hat{h}_{cv}\}^{-1} \sum \sum_{i \ne j} K\{(X_i - X_j)/\hat{h}_{cv}\} \le 5(\sup K)\,(n^2 \hat{h}_{cv})^{-1} ,$$

and more generally,

$$\begin{aligned}
\int \hat{f}(x|\hat{h}_{cv})^2 \, dx &= (n\hat{h}_{cv})^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int K\{(x - X_i)\hat{h}_{cv}^{-1}\} \, K\{(x - X_j)\,\hat{h}_{cv}^{-1}\} \, dx \\
&\ge (n\hat{h}_{cv})^{-2} \sum_{i=1}^{n} \int K\{(x - X_i)\,\hat{h}_{cv}^{-1}\}^2 \, dx \\
&= (n\hat{h}_{cv})^{-1} \int K^2 .
\end{aligned}$$

Therefore, provided $n \geq 20(\sup K)/(\int K^2)$,

$$\begin{aligned}\mathrm{CV}(\hat{h}_{cv}) &= \int \hat{f}(x|\hat{h}_{cv})^2\,dx - 2\{n(n-1)\,\hat{h}_{cv}\}^{-1} \sum\sum_{i\neq j} K\{(X_i - X_j)/\hat{h}_{cv}\} \\ &\geq \tfrac{1}{2}\int \hat{f}(x|\hat{h}_{cv})^2\,dx\,.\end{aligned}$$

Since $\hat{h}_{cv}$ minimizes CV then $\mathrm{CV}(\hat{h}_{cv}) \leq \mathrm{CV}(h_\mu)$, and so

$$\int \hat{f}(x|\hat{h}_{cv})^2\,dx \leq 2\mathrm{CV}(h_\mu)\,,$$

whence

$$\int \{\hat{f}(x|\hat{h}_{cv}) - f(x)\}^2 \leq 4\mathrm{CV}(h_\mu) + 2\int f^2\,,$$

implying that

$$\begin{aligned}q_2 &\leq 4E\Big[\Big\{|\mathrm{CV}(h_\mu)| + \int f^2\Big\}\, I(\hat{h}_{cv} \leq n^{-b})\Big] \\ &\leq 4\Big([E\{\mathrm{CV}(h_\mu)^2\}]^{\frac{1}{2}} + \int f^2\Big)\, P(\hat{h}_{cv} \leq n^{-b})^{\frac{1}{2}}\,.\end{aligned}$$

It is straightforward to prove that $E\{\mathrm{CV}(h_\mu)^2\}$ is bounded, and so

$$q_2 = O\big\{P(\hat{h}_{cv} \leq n^{-b})^{\frac{1}{2}}\big\}\,,$$

which is a sufficient bound for q_2. □

LEMMA 5.2. *Under the conditions of the Theorem, there exists a constant $C > 0$ such that for all $\lambda > 0$,*

$$E\Big[\int \{\hat{f}(x|\hat{h}_{cv}) - f(x)\}^2\,dx\, I(\hat{h}_{cv} > C)\Big] = O(n^{-\lambda})\,.$$

PROOF. The proof is in two steps.

Step (i): *Bound for* $\mathrm{CV}(h_\mu)$. Define $L(u) = \int K(u+v)\,K(v)\,dv$, $W_1 = L$, $W_2 = K$,

$$\mu_k(x) = E[W_k\{(x - X)\,h^{-1}\}]\,, \quad \mu_k = E\{\mu_k(X)\}\,, \quad \nu_k(x) = \mu_k(x) - hf(x)\,,$$

$$\nu_k = E\{\nu_k(X)\}\,, \quad S_k = (n^2h)^{-1} \sum\sum_{i\neq j} [W_k\{(X_i - X_j)\,h^{-1}\} - \mu_k(X_i) - \mu_k(X_j) + \mu_k]\,,$$

$$T_k = (nh)^{-1} \sum_{i=1}^{n} \{\nu_k(X_i) - \nu_k\}\,, \quad U = n^{-1} \sum_{i=1}^{n} \Big\{f(X_i) - \int f^2\Big\}\,.$$

In this notation,

$$\int \hat{f}_h(x)^2\,dx = (nh)^{-1}\int K^2 + (1-n^{-1})\,h^{-1}\,\mu_1 + S_1 + 2(1-n^{-1})\,(T_1+U)\,,$$

$$(n^2h)^{-1}\sum\sum_{i\neq j} K\{(X_i-X_j)\,h^{-1}\} = (1-n^{-1})\,h^{-1}\,\mu_2 + S_2 + 2(1-n^{-1})\,(T_2+U)\,,$$

whence

$$\begin{aligned} D(h) &= \mathrm{CV}(h) + \int f^2 - M(h) \\ &= S_1 - 2(1-n^{-1})^{-1}\,S_2 + 2(1-n^{-1})\,T_1 - 4T_2 - 2(1+n^{-1})\,U\,. \end{aligned}$$

Rosenthal's inequality (Hall and Heyde 1980, p.23) may be used to prove that for all integers $p \geq 1$,

$$\begin{aligned} &E(S_k^{2p}) = O\{(n^2h)^{-p}\}\,, \quad E(T_k^{2p}) = O\{(nh^{-4})^{-p}\}\,, \\ &E(U^{2p}) = O(n^{-p}) \end{aligned} \tag{5.3}$$

uniformly in $0 < h \leq C_1$, for any $C_1 > 0$. Therefore, by Markov's inequality,

$$P\{|D(h_\mu)| > C_2\} = O(n^{-\lambda})$$

for all $C_2 > 0$ and $\lambda > 0$. Thus,

$$\begin{aligned} P\Big\{\mathrm{CV}(h_\mu) > -\tfrac{1}{2}\int f^2\Big\} &= P\Big\{D(h_\mu) > \tfrac{1}{2}\int f^2 - M(h_\mu)\Big\} \\ &= O(n^{-\lambda})\,. \end{aligned} \tag{5.4}$$

Step (ii): *Completion.* Observe that $|\mathrm{CV}(h)| \leq C_2 h^{-1}$, where $C_2 = \int K^2 + 2\sup K$. Therefore, $|\mathrm{CV}(h)| < C_2\,C_3^{-1}$ uniformly in $h > C_3$. Hence, if C_3 is chosen so large that $C_2\,C_3^{-1} \leq \frac{1}{2}\int f^2$,

$$\begin{aligned} P(\hat{h}_{cv} > C_3) &\leq P\{\mathrm{CV}(h_\mu) > -C_2\,C_3^{-1}\} \\ &\leq P\Big\{\mathrm{CV}(h_\mu) > -\tfrac{1}{2}\int f^2\Big\} = O(n^{-\lambda}) \end{aligned} \tag{5.5}$$

for all $\lambda > 0$, using (5.4). Consequently,

$$\begin{aligned} &E\Big[\int\{\hat{f}(x|\hat{h}_{cv}) - f(x)\}^2\,dx\,I(\hat{h}_{cv} > C_3)\Big] \\ &\qquad \leq \Big(C_3^{-1}\int K^2 + \int f^2\Big)\,P(\hat{h}_{cv} > C_3) = O(n^{-\lambda}) \end{aligned}$$

for all $\lambda > 0$, as had to be shown. □

LEMMA 5.3. *Under the conditions of the Theorem, and for each* $\epsilon, \lambda > 0$,

$$P\big(|\hat{h}_{cv} - h_\mu| > n^{-\frac{3}{10}+\epsilon}\big) = O(n^{-\lambda}).$$

PROOF. Since f is bounded then so is the density of $|X_1 - X_2|$, and hence

$$P(|X_1 - X_2| \le h) = O(h)$$

as $h \to 0$. Therefore,

$$P\Big[\sum\sum_{i<j} K\{(X_i - X_j)\,h^{-1}\} > 0\,,\ \text{for some } h \le n^{-b}\Big] \le n^2\,P(|X_1 - X_2| \le n^{-b}) = O(n^{-(b-2)}).$$

It follows that

$$P\Big\{\inf_{h<n^{-b}} \mathrm{CV}(h) < 0\Big\} = O(n^{-(b-2)}).$$

We know from (5.4) that $P\{\mathrm{CV}(h_\mu) \ge 0\} = O(n^{-\lambda})$ for all $\lambda > 0$, and so $P(\hat{h}_{cv} \le n^{-b}) = O(n^{-(b-2)})$. Also, by (5.5), there exists a constant $C_1 > 0$ such that $P(\hat{h}_{cv} > C_1) = O(n^{-\lambda})$ for all $\lambda > 0$. Hence, it suffices to show that for all $b, \epsilon, \lambda > 0$,

$$P(|\hat{h}_{cv} - h_\mu| > n^{-\frac{3}{10}+\epsilon},\ n^{-b} \le \hat{h}_{cv} \le C_1) = O(n^{-\lambda}). \tag{5.6}$$

Our proof of this result contains two steps, of which the first parallels Step (i) in the proof of Lemma 5.2.

Step (i): *Bound for* $\mathrm{CV}'(h)$. Let L, $\mu_k(x)$, μ_k, $\nu_k(x)$, ν_k, S_k, T_k be as defined in Step (i) of the proof of Lemma 5.2, except that we re-define $W_1(u) = L(u) + uL'(u)$ and $W_2(u) = K(u) + uK'(u)$. In this notation

$$-(d/dh)\int \hat{f}_h(x)^2\,dx = (nh^2)^{-1}\int K^2 + (1-n^{-1})h^{-2}\,\mu_1 + h^{-1}\,S_1 + 2(1-n^{-1})h^{-1}\,T_1\,,$$

$$-(d/dh)\,(n^2h)^{-1}\sum\sum_{i\ne j} K\{(X_i - X_j)\,h^{-1}\} = (1-n^{-1})\,h^{-2}\,\mu_2 + h^{-1}\,S_2 + 2(1-n^{-1})\,h^{-1}\,T_2\,.$$

It follows that

$$D'(h) = \mathrm{CV}'(h) - M'(h)$$
$$= -h^{-1}\{S_1 - 2(1-n^{-1})\,S_2 + 2(1-n^{-1})\,T_1 - 4T_2\}\,.$$

Rosenthal's inequality may be used to prove (5.3) once again, in the present notation, whence it follows that

$$E\{D'(h)^{2p}\}\{(nh^3)^{-1} + h^2\}^{-2p} = O[\{h^3 \wedge (n^{-1}\,h^{-2})\}^p]\,. \tag{5.7}$$

Under the conditions of the theorem, there exists a constant $C_2 > 0$ such that

$$M'(h) \geq C_2|h - h_\mu|\,\{(nh^3)^{-1} + h^2\} \qquad \text{for } 0 < h \leq h_\mu\,,$$
$$M'(h) \leq -C_2|h - h_\mu|\,\{(nh^3)^{-1} + h^2\} \qquad \text{for } h_\mu < h \leq C_1\,.$$

Hence,

$$\mathrm{CV}'(h) \geq C_2|h - h_\mu|\,\{(nh^3)^{-1} + h^2\} + D'(h) \qquad \text{for } 0 < h \leq h_\mu\,, \tag{5.8}$$
$$\mathrm{CV}'(h) \leq -C_2|h - h_\mu|\,\{(nh^3)^{-1} + h^2\} + D'(h) \qquad \text{for } h_\mu < h \leq C_1\,. \tag{5.9}$$

Let $h_1 = C_3\,n^{-\frac{1}{5}}$ denote the minimizer of $(nh^3)^{-1} + h^2$, and put

$$\eta_n = \tfrac{1}{2}\,C_2\,n^{-\frac{3}{10}}\{(nh_1^3)^{-1} + h_1^2\}\,.$$

Then by (5.8), and for $0 < h < h_\mu - n^{-\frac{3}{10}+\epsilon}$,

$$P\{\mathrm{CV}'(h) < \eta_n\} \leq 2^{2p}\left[\tfrac{1}{2}\,C_2|h - h_\mu|\,\{(nh^3)^{-1} + h^2\}\right]^{-2p} E\{D'(h)^{2p}\}\,,$$

whence we may prove from (5.7) that for all $\epsilon, \lambda > 0$,

$$\sup_{0<h\leq h_\mu - n^{-(3/10)+\epsilon}} P\{\mathrm{CV}'(h) < \eta_n\} = O(n^{-\lambda})\,. \tag{5.10}$$

Similarly, by (5.7) and (5.9),

$$\sup_{h_\mu + n^{-(3/10)+\epsilon} \leq h \leq C_1} P\{\mathrm{CV}'(h) > -\eta_n\} = O(n^{-\lambda})\,. \tag{5.11}$$

Step (ii): *Completion.* Let $m > 0$ be fixed but arbitrarily large, put $t_i = i\,n^{-m}$ for $i \geq 1$, and given $h > 0$ let $t(h)$ denote a value of t_i which minimizes $|h - t_i|$ over $i \geq 1$. Given $b, c > 0$ we may choose m so large that

$$\sup_{n^{-b} \leq h \leq C_1} |\mathrm{CV}'(h) - \mathrm{CV}'\{t(h)\}| \leq n^{-c}\,, \tag{5.12}$$

all $n \geq n_0$. By choosing c sufficiently large we may ensure that for $n \geq 2$, $n^{-c} < \eta_n$. Hence for some $m > 0$, and writing

$$H_1 = [n^{-b}, h_\mu - n^{-\frac{3}{10}+\epsilon}]\,, \qquad H_2 = [h_\mu + n^{-\frac{3}{10}+\epsilon}, C_1]\,,$$

we have

$$\begin{aligned}
&P\{|\hat{h}_{cv} - h_\mu| > n^{-\frac{3}{10}+\epsilon},\ n^{-b} \leq \hat{h}_{cv} \leq C_1\} \\
&\qquad \leq P\{\mathrm{CV}'(h) = 0\,,\ \text{some } h \in H_1\,;\quad \text{or}\quad \mathrm{CV}'(h) = 0\,,\ \text{some } h \in H_2\} \\
&\qquad \leq P\Big\{ \inf_{h \in H_1} \mathrm{CV}'(h) \leq 0 \ \text{ or } \ \sup_{h \in H_2} \mathrm{CV}'(h) \geq 0 \Big\} \\
&\qquad \leq P\Big[\inf_{h \in H_1} \mathrm{CV}'\{t(h)\} \leq \eta_n \ \text{ or } \ \sup_{h \in H_2} \mathrm{CV}'\{t(h)\} \geq -\eta_n \Big] \\
&\qquad \leq C_1\, n^m \Big[\sup_{h \in H_1} P\{\mathrm{CV}'(h) \leq \eta_n\} + \sup_{h \in H_2} P\{\mathrm{CV}'(h) \geq -\eta_n\} \Big] \\
&\qquad = O(n^{-\lambda})
\end{aligned}$$

for all $\lambda > 0$, the last line following from (5.10) and (5.11). This proves (5.6). □

A similar proof may be used to derive the following result.

LEMMA 5.4. *Under the conditions of the Theorem, and for each* $\epsilon, \lambda > 0$,

$$\begin{aligned}
P(|\hat{h} - h_\mu| > n^{-\frac{3}{10}+\epsilon}) &= O(n^{-\lambda}) \quad \textit{if } \hat{h} = \hat{h}_\nu \textit{ or } \hat{h}_{\dot{\nu}}\,, \\
P(|\hat{h} - h_\mu| > n^{-\frac{1}{2}+\epsilon}) &= O(n^{-\lambda}) \quad \textit{if } \hat{h} = \hat{h}_{pi,1} \textit{ or } \hat{h}_{pi,2}\,.
\end{aligned}$$

LEMMA 5.5. *Under the conditions of the Theorem,*

$$E\{\Delta(h_\mu)^2\} = O(n^{-\frac{8}{5}}), \qquad E\{\Delta'(h_\mu)^2\} = O(n^{-\frac{6}{5}}),$$
$$E[\{\Delta''(h_\mu) - M''(h_\mu)\}^2] = O(n^{-1}), \tag{5.13}$$

and for some $\eta > 0$ *and all* $0 < \epsilon < 1$,

$$E\Big[\sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} \{\Delta''(h) - \Delta''(h_\mu)\}^2\Big] = O(n^{-\frac{4}{5}-\eta\epsilon}). \tag{5.14}$$

PROOF. Results (5.13) are relatively straightforward to derive, and so we confine attention to proving (5.14). The argument is similar in many respects to that used to derive Lemma 5.3, and so we give only a sketch.

Let $t_i, i \ge 1$, and $t(h)$ denote the quantities defined in Step (ii) of the proof of Lemma 5.3, and choose m so large that, instead of (5.12),

$$S_1 = \sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} [\Delta''(h) - \Delta''\{t(h)\}]^2 \le n^{-10}$$

for all $n \ge n_0$. Put

$$S_2 = \sup_{i\ge 1:|t_i-h_\mu|\le n^{-(1/5)-\epsilon}} \{\Delta''(t_i) - \Delta''(h_\mu)\}^2.$$

Since the left-hand side of (5.14) is dominated by $4E(S_1 + S_2)$ then it suffices to prove that for some $\eta > 0$, and all $0 < \epsilon < 1$, $E(S_2) = O(n^{-\frac{4}{5}-\eta\epsilon})$.

The number of i's such that $|t_i - h_\mu| \le n^{-\frac{1}{5}-\epsilon}$ is of order n^m. Hence, for any integer $p \ge 1$,

$$E(S_2) \le \Big[\sum_{i\ge 1:|t_i-h_\mu|\le n^{-(1/5)-\epsilon}} E\{\Delta''(t_i) - \Delta''(h_\mu)\}^{2p}\Big]^{1/p}$$
$$= O(n^{m/p}) \sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} [E\{\Delta''(t_i) - \Delta''(h_\mu)\}^{2p}]^{1/p}.$$

Therefore, it suffices to show that for some $\eta > 0$ and any $p \ge 1$ and $0 < \epsilon \le 1$,

$$\sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} E\{\Delta''(h) - \Delta''(h_\mu)\}^{2p} = O(n^{-p(\frac{4}{5}+\eta\epsilon)}). \tag{5.15}$$

Arguing as in Step (i) of the proof of Lemma 5.3 we may show that

$$\sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} E\{\Delta''(h) - M''(h)\}^{2p} = O(n^{-\frac{2p}{5}}). \tag{5.16}$$

Under condition (3.16) it may be proved that for some $\eta > 0$ and all $0 < \epsilon \le 1$,

$$\sup_{h:|h-h_\mu|\le n^{-(1/5)-\epsilon}} |M''(h) - M''(h_\mu)| = O(n^{-\frac{2}{5}-\eta\epsilon}). \tag{5.17}$$

The desired result (5.15) follows from (5.16) and (5.17). □

LEMMA 5.6. *Under the conditions of the Theorem,*

$$n^{2\alpha}\, E\{(\hat{h} - \hat{h}_\nu)^2\, I(|\hat{h} - \hat{h}_\nu| \le n^{-a})\} \to \sigma^2 \tag{5.18}$$

for $\hat{h} = \hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi,1}$ *and* $\hat{h}_{pi,2}$, *and*

$$n^{\frac{3}{5}}\, E\{(\hat{h}_\nu - h_\mu)^2\, I(|\hat{h}_\nu - h_\mu| \le n^{-\frac{3}{10}+\epsilon})\} \to \sigma^2_{\mu,\nu}. \tag{5.19}$$

The cases of $\hat{h}_\nu$, $\hat{h}_{cv}$ may be deduced from Hall and Marron (1987a); $\hat{h}_{\hat{\nu}}$ may be obtained from Hall and Johnstone (1992); and $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$, from Hall and Marron (1987b) and Hall *et al.* (1991).

Now we conclude the proof of the theorem. First we decompose $\mu(\hat{h})$ into four parts. Let α, a be as defined in the first paragraph of the proof, and let $b \gg a$ denote a fixed but arbitrarily large positive constant. Fix $C > 0$, and write $\mathcal{E}_1, \ldots, \mathcal{E}_4$ for the respective events $\hat{h} \le n^{-b}$, $\hat{h} > C$, $\{|\hat{h} - h_\mu| > n^{-a}\} \cap \{n^{-b} < \hat{h} \le C\}$, $|\hat{h} - h_\mu| \le n^{-a}$, respectively. Define

$$\nu_j = E\{\Delta(\hat{h})\, I(\mathcal{E}_j)\}, \qquad 1 \le j \le 4.$$

We claim for all $\lambda > 0$,

$$\nu_1 + \nu_2 + \nu_3 = O(n^{-\lambda}), \tag{5.20}$$

which implies that the asymptotic properties of $\nu(\hat{h}) = \sum \nu_j$ are determined by those of ν_4.

In the case $\hat{h} = \hat{h}_{cv}$ we deduce from Lemmas 5.1–5.3 that

$$\nu_1 + \nu_2 = O(n^{-\lambda}) \tag{5.21}$$

for all $\lambda > 0$. Since $\inf_h \Delta(h) \le \int f^2$ then when $\hat{h} = \hat{h}_\nu$,

$$\nu_1 + \nu_2 \le \Big(\int f^2 \Big) P(\hat{h} \le n^{-a}, \hat{h} > C) = O(n^{-\lambda}),$$

for all $\lambda > 0$, the last identity following from Lemma 5.4. In the cases $\hat{h} = \hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$ and $\hat{h}_{pi,2}$, we note that either (in view of (3.9)) $P(\hat{h} \ge n^{-r}) = 1$ for some $r > 0$, and without loss of generality $b > r$; or (in view of (3.10) and (5.2)),

$$\begin{aligned} \nu_1 &\le 2\Big(\int K^2 + \int f^2 \Big) [E\{\kappa_1(n|\hat{J}_2|)^{-\frac{1}{5}}\}^{-2} P(\hat{h} \le n^{-b})]^{\frac{1}{2}} \\ &= O\{n^{\frac{1}{5}} P(\hat{h} \le n^{-b})^{\frac{1}{2}}\} = O(n^{-\lambda}) \end{aligned}$$

for all $\lambda > 0$, the last identity following from Lemma 5.4. Thus, for those three variants of $\hat{h}$, $\nu_1 = O(n^{-\lambda})$. Similarly, for $\hat{h} = \hat{h}_{\hat{\nu}}$, $\hat{h}_{pi,1}$ or $\hat{h}_{pi,2}$,

$$\nu_2 \le 2\Big(C^{-1} \int K^2 + \int f^2 \Big) P(\hat{h} > C) = O(n^{-\lambda})$$

for all $\lambda > 0$. Therefore, (5.21) holds for $\hat{h} = \hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi}$.

By Lemmas 5.3 and 5.4, when $\hat{h} = \hat{h}_\nu$, $\hat{h}_{\hat{\nu}}$, $\hat{h}_{cv}$, $\hat{h}_{pi,1}$ or $\hat{h}_{pi,2}$,

$$\nu_3 \le 2\Big(n^b \int K^2 + \int f^2 \Big) P(|\hat{h} - h_\mu| > n^{-a}) = O(n^{-\lambda})$$

for all $\lambda > 0$. This completes the proof of (5.20).

Next we develop a Taylor series expansion of ν_4, of the form

$$\nu_4 = M(h_\mu) + \tfrac{1}{2}(\sigma^2 - \sigma^2_{\mu,\nu})\, c_3\, n^{-1} + o(n^{-1}). \tag{5.22}$$

Result (3.19) follows from (5.20) and (5.22).

Observe from Lemma 5.5 that for some $\eta > 0$,

$$\begin{aligned} &\big| \nu_4 - E\big[\{\Delta(h_\mu) + (\hat{h} - h_\mu)\Delta'(h_\mu) + \tfrac{1}{2}(\hat{h} - h_\mu)^2 \Delta''(h_\mu)\}\, I(\mathcal{E}_4) \big] \big| \\ &\qquad \le n^{-2a} \Big(E\Big[\sup_{h:|h-h_\mu| \le n^{-a}} \{\Delta''(h) - \Delta''(h_\mu)\}^2 \Big] \Big)^{\frac{1}{2}} \\ &\qquad = O(n^{-2a - \frac{2}{5} - \eta a}) = o(n^{-2\alpha - \frac{2}{5}}), \end{aligned}$$

provided $\epsilon > 0$ is chosen sufficiently small. By Lemmas 5.3, 5.4 and 5.5,

$$E\{\Delta(h_\mu)\, I(\tilde{\mathcal{E}}_4)\} \le [E\{\Delta(h_\mu)^2\}\, P(|\hat{h} - h_\mu| > n^{-a})]^{\frac{1}{2}} = O(n^{-\lambda})$$

for all $\lambda > 0$; and for some $\eta > 0$,

$$\begin{aligned} E[(\hat{h} - h_\mu)^2 |\Delta''(h_\mu) - M''(h_\mu)|\, I(\mathcal{E}_4)] &\le n^{-2a}\, [E\{\Delta''(h_\mu) - M''(h_\mu)\}^2]^{\frac{1}{2}} \\ &= O(n^{-2a-\frac{2}{5}-\eta a}) = o(n^{-2\alpha-\frac{2}{5}}), \end{aligned}$$

provided $\epsilon > 0$ is sufficiently small. Therefore,

$$\begin{aligned} |\nu_4 - M(h_\mu) - E\{(\hat{h} - h_\mu)\, \Delta'(h_\mu)\, I(\mathcal{E}_4)\} - \tfrac{1}{2}\, E\{(\hat{h} - h_\mu)^2\, I(\mathcal{E}_4)\}\, M''(h_\mu)| \\ = o(n^{-2\alpha-\frac{2}{5}}). \end{aligned} \tag{5.23}$$

Put $a_\nu = \frac{3}{10} - \epsilon$, so that a_ν is the value of a for $\hat{h}_\nu$. Define $\mathcal{E}_5 = \mathcal{E}_4 \cap \{|\hat{h}_\nu - h_\mu| \le n^{-a_\nu}\}$ and $\mathcal{E}_6 = \mathcal{E}_4 \backslash \mathcal{E}_5$. By Lemma 5.4, $P(\mathcal{E}_6) = O(n^{-\lambda})$ for all $\lambda > 0$, and so

$$\begin{aligned} E\{|(\hat{h} - h_\mu)\, \Delta'(h_\mu)|\, I(\mathcal{E}_6)\} &\le n^{-a}\, [E\{\Delta'(h_\mu)^2\}\, P(\mathcal{E}_6)]^{\frac{1}{2}} \\ &= O(n^{-\lambda}) \end{aligned}$$

for all $\lambda > 0$. By Lemma 5.5,

$$\begin{aligned} &E\big[|(\hat{h} - h_\mu)\, \{\Delta'(h_\mu) - (h_\mu - \hat{h}_\nu)\, \Delta''(h_\mu)\}|\, I(\mathcal{E}_5)\big] \\ &\qquad \le n^{-a-a_\nu} \Big(E\Big[\sup_{h:|h-h_\mu|\le n^{-a_\nu}} \{\Delta''(h) - \Delta''(h_\mu)\}^2\Big]\Big)^{\frac{1}{2}} \\ &\qquad = O\big(n^{-a-\frac{7}{10}+\epsilon(1+\eta)-\eta\frac{1}{10}}\big) = o\big(n^{-\alpha-\frac{7}{10}}\big), \\ &E\big[|(\hat{h} - h_\mu)(h_\mu - \hat{h}_\nu)\, \{\Delta''(h_\mu) - M''(h_\mu)\}|\, I(\mathcal{E}_5)\big] \\ &\qquad \le n^{-a-a_\nu}\, (E\{\Delta''(h_\mu) - M''(h_\mu)\}^2)^{\frac{1}{2}} = o(n^{-\alpha-\frac{7}{10}}), \end{aligned}$$

provided $\epsilon > 0$ (in the definition of $a = \alpha - \epsilon$) is chosen sufficiently small. Hence,

$$E\{(\hat{h} - h_\mu)\, \Delta'(h_\mu)\, I(\mathcal{E}_4)\} = E\{(\hat{h} - h_\mu)(h_\mu - \hat{h}_\nu)\, I(\mathcal{E}_5)\}\, M''(h_\mu) + o(n^{-\alpha-\frac{7}{10}}).$$

Now,

$$2(\hat{h}-h_\mu)(h_\mu-\hat{h}_\nu)=(\hat{h}-\hat{h}_\nu)^2-(\hat{h}-h_\mu)^2-(\hat{h}_\nu-h_\mu)^2\,,$$

$$\begin{aligned}E\{(\hat{h}-\hat{h}_\nu)^2\,|I(\mathcal{E}_5)-I(|\hat{h}-\hat{h}_\nu|\le n^{-a_\nu})|\}&\\ \le 4n^{-2a_\nu}\{P(|\hat{h}-h_\mu|>\tfrac{1}{2}\,n^{-a_\nu})+P(|\hat{h}_\nu-h_\mu|>\tfrac{1}{2}\,n^{-a_\nu})\}&\\ =O(n^{-\lambda})\,,&\end{aligned}$$

$$E\{(\hat{h}-h_\mu)^2|I(\mathcal{E}_5)-I(|\hat{h}-h_\mu|\le n^{-a_\nu})|\}=O(n^{-\lambda})\,,$$

$$E\{(\hat{h}-h_\mu)^2|I(\mathcal{E}_5)-I(|\hat{h}_\nu-h_\mu|\le n^{-a_\nu})|\}=O(n^{-\lambda})\,,$$

for all $\lambda>0$. Therefore,

$$\begin{aligned}2E\{(\hat{h}-h_\mu)\,\Delta'(h_\mu)\,I(\mathcal{E}_4)\}=[&E\{(\hat{h}-\hat{h}_\nu)^2\,I(|\hat{h}-\hat{h}_\nu|\le n^{-a})\}\\ &-E\{(\hat{h}-h_\mu)^2\,I(|\hat{h}-h_\mu|\le n^{-a})\}\\ &-E\{(\hat{h}_\nu-h_\mu)^2\,I(|\hat{h}_\nu-h_\mu|\le n^{-a})\}]\,M''(h_\mu)\\ &+o(n^{-\alpha-\frac{7}{10}})\,.\end{aligned}\tag{5.24}$$

Combining (5.23) and (5.24) we deduce that

$$\begin{aligned}\nu_4=M(h_\mu)+\tfrac{1}{2}\,[&E\{(\hat{h}-\hat{h}_\nu)^2\,I(|\hat{h}-\hat{h}_\nu|\le n^{-a})\}\\ &-E\{(\hat{h}_\nu-h_\mu)^2\,I(|\hat{h}_\nu-h_\mu|\le n^{-a})\}]\,M''(h_\mu)+o(n^{-1})\,.\end{aligned}$$

The desired result (5.22) follows from this formula and Lemma 5.6. □

6. References

ALDERSHOF, B.K. (1991). Estimation on integrated squared density derivatives. PhD Dissertation, North Carolina Institute of Statistics, Mimeo Series No. 2053.

CAO-ABAD, R., CUEVAS, A. AND GONZALEZ-MANTEIGA, W. (1992). A comparative study of several smoothing methods in density estimation. To appear in *Computational Statistics and Data Analysis.*

FAN, J. AND MARRON, J.S. (1992). Best possible constant for bandwidth selection. To appear in *Ann. Statist.*

HALL, P. AND HEYDE, C.C. (1980). Martingale Limit Theory and its Application. Academic Press, New York.

HALL, P. AND JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54**, 475–530.

HALL, P. AND MARRON, J.S. (1987a). Extent to which least-squares cross-validationion minimises integrated squared error in nonparametric density estimation. *Prob. Th. Rel. Fields* **74**, 567–581.

HALL, P. AND MARRON, J.S. (1987b). Estimation of integrated squared density derivatives. *Statist. Prob. Lett.* **6**, 109–115

HALL, P. AND MARRON, J.S. (1987c). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15**, 163–181.

HALL, P. AND MARRON, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Prob. Th. Rel. Fields* **90**, 149–173.

HALL, P., SHEATHER, S.J., JONES, M.C. AND MARRON, J.S. (1991). On optimal data-based bandwidth selection in density estimation. *Biometrika*, **78**, 263–271.

JONES, M.C. (1991). The role of ISE and MISE in density estimation. *Statist. Probab. Lett.* **12**, 51–56.

JONES, M.C., MARRON, J.S. AND PARK, B. (1991). A simple root-n bandwidth selector. *Ann. Statist.* **19**, 1919–1932.

JONES, M.C., MARRON, J.S. AND SHEATHER, S.J. (1992). Progress in data based bandwidth selection for kernel density estimation. Unpublished manuscript.

JONES, M.C. AND SHEATHER, S.J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11**, 511 – 514.

MAMMEN, E. (1990). A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Lett.* **9**, 23–25.

MARRON, J.S. (1988). Automatic smoothing parameter selection: a survey. *Empir. Econ.* **13**, 187–208.

MARRON, J.S. AND WAND, M. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.

PARK, B.U. AND MARRON, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85**, 66–72.

PARK, B.U. AND TURLACH, B.A. (1992). Practical performance of several data driven bandwidth selectors (with discussion). To appear in *Comp. Statist.*.

SHEATHER, S.J. AND JONES, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53**, 683–690.

SILVERMAN, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Captions

Figure 1. Comparison of the empirical performance of $\hat{h}_{HJS1}$ and $\hat{h}_{PMPI}$, measured by μ and ν, together with the asymptotically leading term of $\nu(\hat{h}_\nu)$ and the empirical value of $\nu(\hat{h}_\nu)$, for the densities #1 to #15 in Marron and Wand (1992).

Figure 1a. Sample size $n = 100$.

Figure 1b. Sample size $n = 1000$.

Figure 2. Comparison of the empirical performance of $\hat{h}_{HJTT}$ and $\hat{h}_{PMPI}$, measured by μ and ν, together with the asymptotically leading term of $\nu(\hat{h}_\nu)$ and the empirical value of $\nu(\hat{h}_\nu)$, for the densities #1 to #15 in Marron and Wand (1992).

Figure 2a. Sample size $n = 100$.

Figure 2b. Sample size $n = 1000$.

Figure 1a
n = 100

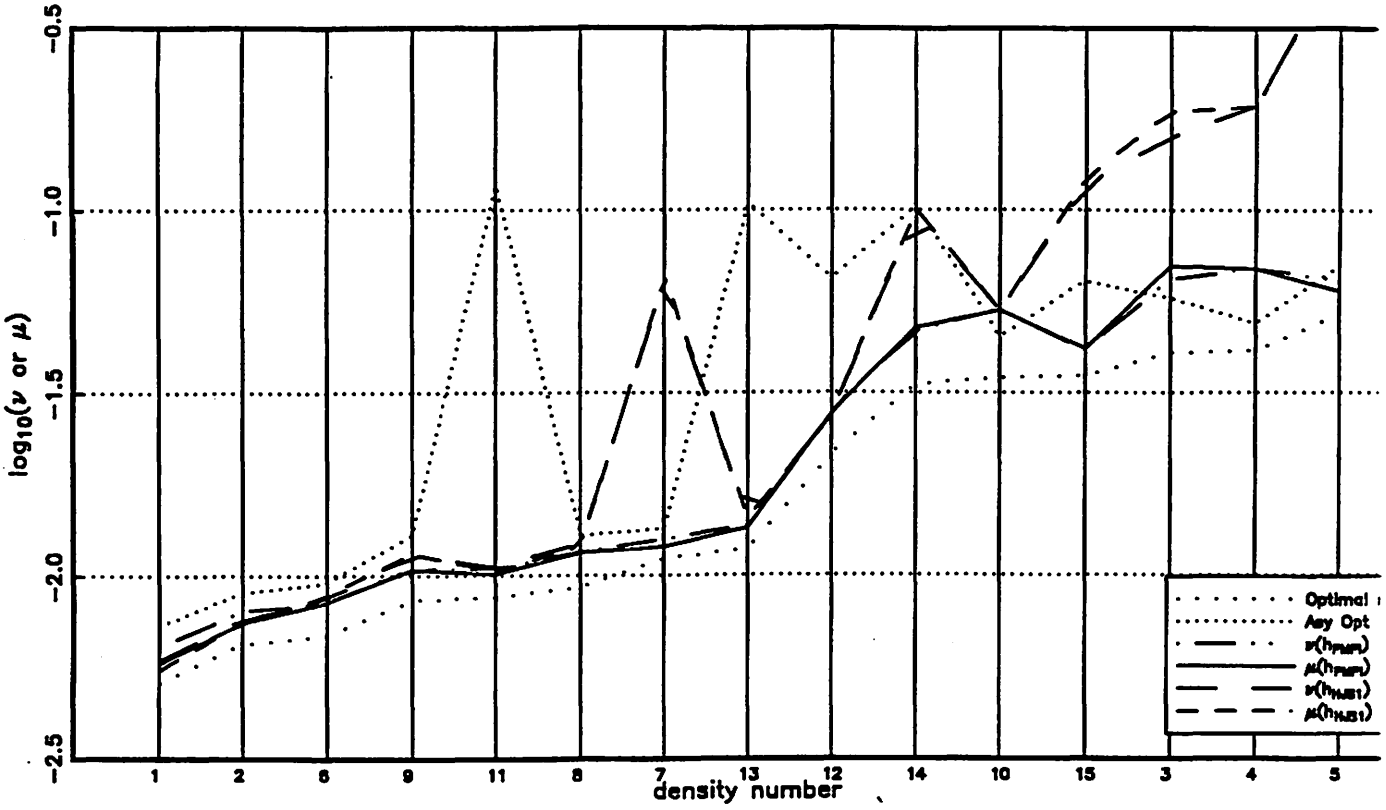

Figure 1b
n = 1000

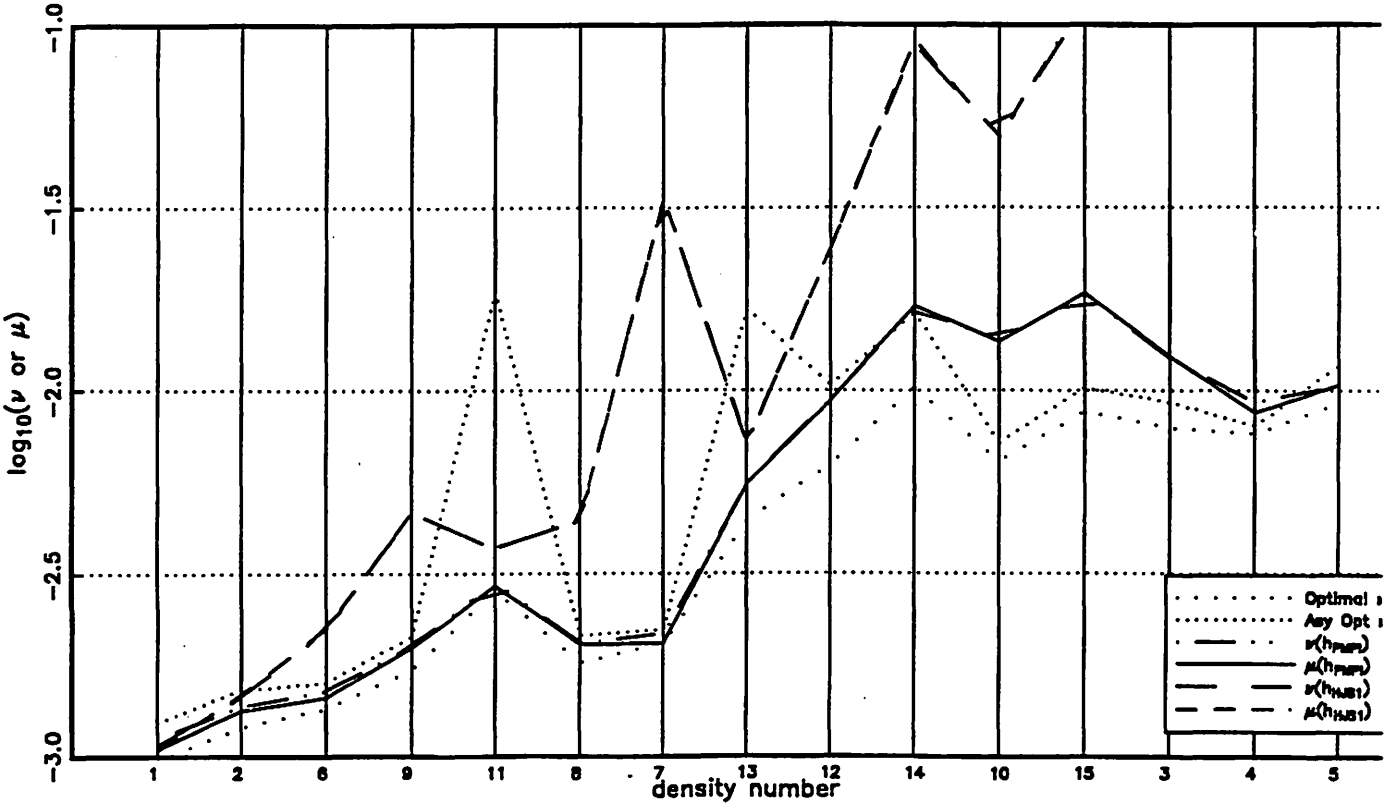

Figure 2a
n = 100

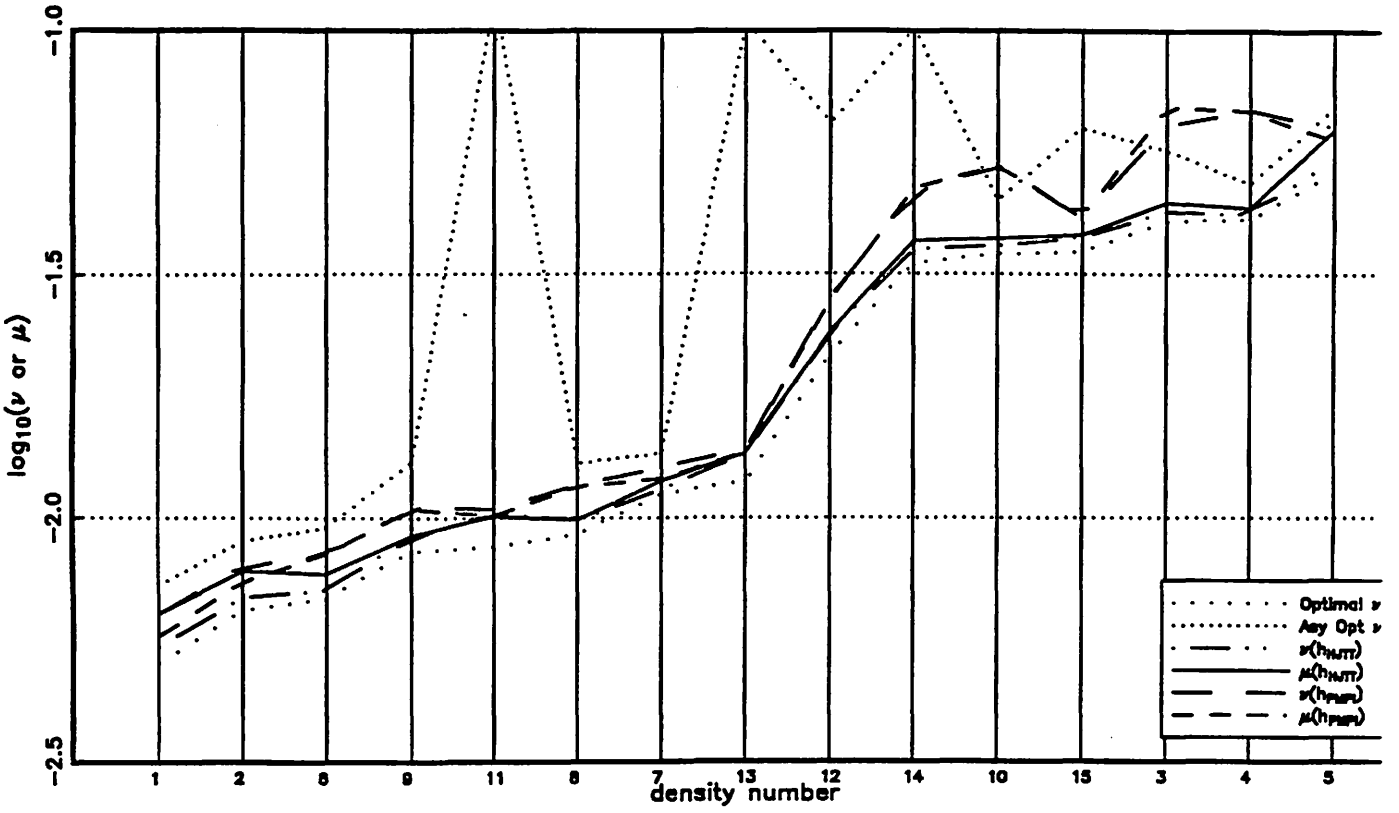

Figure 2b
n = 1000